Factor Scoring Methods Affected by Response Shift in Patient-Reported Outcomes

A Thesis Submitted to the

College of Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the degree of Master of Science

in the Department of Mathematics and Statistics

University of Saskatchewan

Saskatoon

By

Courtney Kendall

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis. Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics

Room 142 McLean Hall

106 Wiggins Road

University of Saskatchewan

Saskatoon, Saskatchewan

Canada

S7N 5E6

Abstract

Objective: Patient-reported outcomes (PROs) are measures collected from a patient to determine how he/she feels or functions in regards to a health condition. Longitudinal PROs, which are collected at multiple occasions from the same individual, may be affected by response shift (RS). RS is a change in a person's self-evaluation of a target construct. Latent variable models (LVMs) are statistical models that relate observed variables to latent variables (LV). LVMs are used to analyze PROs and detect RS. LVs are random variables whose realizations are not observable. Factor scores are estimates of LVs for each individual and can be estimated from parameter estimates of LVMs. Factor scoring methods to estimate factor scores include: Thurstone, Bartlett, and sum scores. This simulation study examines the effects of RS on factor scores used to test for change in the LV means and recommend a factor scoring method least affected by RS.

Methods: Data from two time points were fit to three confirmatory factor analysis (CFA) models. CFA models are a type of LVM. Each CFA model had different sets of parameters that were invariant over time. The unconstrained (Uncon) CFA model had no invariant parameters, the constrained (Con) model had all the parameters invariant, and the partially constrained (Pcon) model had some of the parameters invariant over time. Factor scores were estimated and tested for change over time via paired t-test. The Type I error, power, and factor loading (the regression coefficient between an observed and LV) and factor score bias were estimated to determine if RS influenced the test of change over time and factor score estimation.

Results: The results depended on the true LV mean. The Type I error and power were similar for all factor scoring methods and CFA models when the LV mean was 0 at time 1. For LV mean of 0.5 at time 1 the Type I error and power increased as RS increased for all factor scores except for scores estimated from the Uncon model and Bartlett method. The biases of the factor loadings

were unaffected by RS when estimated from an Uncon model. The factor scores estimated from the Uncon model and the Bartlett and sum scores method had the smallest factor score biases. **Conclusion:** The factor scores estimated from the Uncon model and the Bartlett method was least affected by RS and performed best in all measures of Type I error, statistical power, factor loading and factor score bias. Estimating factor scores from PROs data that ignores RS may result in erroneous (or biased) estimates.

Acknowledgements

This research is supported by a Canadian Institutes of Health Research (CIHR) Operating Grant to a research team led by Dr. Lisa Lix, the University of Saskatchewan, the School of Public Health, the Department of Mathematics and Statistics, and my graduate stipend. I would like to thank my supervisors; Dr. Lisa Lix and Dr. Juxin Liu for all their help, experience, and advice on this research. Also, thank you to my advisory committee, Dr. Chris Soteros, Dr. Mik Bickis, and Dr. Longhai Li for their suggestions and advice on my thesis work. Thank you to Dr. Tolu Sajobi for his intellectual and technical support of my simulation study. Thank you to the Department of Mathematics and Statistics and the School of Public Health for their support enabling me to conduct and finish this research. Lastly, I would like to dedicate this thesis and thank my husband for all his support and love.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

Patient-reported outcome (PRO) measures are information collected directly from a patient to determine how he/she feels or functions in relation to a health condition and/or treatment. PROs are increasing in popularity in clinical and epidemiologic studies because they can provide insights into patients' perceptions about their own health. They are used to measure health-related quality of life (HRQOL), satisfaction with treatment, functional ability, and disease symptoms. Many PRO data come from instruments (i.e. questionnaires) that have been shown to be credible, reliable and valid. HRQOL measures are typical of many PROs; examples include the Short Form Questionnaire (SF-36), Chronic Respiratory Disease Questionnaire (CRQ), and Quality of Life Scale (QOLS)[5, 34, 40]. In studies about HRQOL, data are often collected on multiple domains, such as physical function, social health, and emotional health. For example, the SF-36 has eight domains; vitality, physical functioning, bodily pain, general health perceptions, physical role functioning, emotional role functioning, social function and mental health. While PRO measures are important for assessment in patient-focused healthcare systems, there are challenges associated with their development and analysis. Longitudinal PRO studies, which are studies that collect PRO measures from patients repeatedly through time, may be affected by response shift (RS) [31].

RS is defined as changes in a person's self-evaluation of a target outcome such as HRQOL. This change results from: changes in internal standards or recalibration of the measurement scale, changes in the definition or conceptualization of the construct, and/or changes in values or prioritization of domains within the construct[35]. RS occurs over time and

can affect the interpretation of change in measures of a target construct collected over time[48]. If RS is present in a set of data, conventional statistical methods may not be able to detect true change in measures. It has been theorized that RS occurs when an individual experiences a significant health event, called a catalyst, such as a stroke, cancer treatment, or new disease diagnosis[51]. Diseases that may have significant long-term health effects, such as chronic diseases, have also been investigated as catalysts for RS.

Latent variable models are widely used to analyze PRO measures[47]. Latent variables are random variables whose realizations are not observable. In contrast, manifest variables are those for which the realizations are observable. A latent variable model is a statistical model that relates manifest (i.e. observed) variables to latent variables. Let p be the number of observed variables, N the number of individuals, and M the number of latent variables. Then a linear equation to define the relationship between manifest and latent variables is:

$$\mathbf{Y}_i = \boldsymbol{\tau} + \boldsymbol{\Gamma}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \tag{1.1}$$

where $\mathbf{Y}_i$ (p x 1) is the vector of observed variables for $i = 1,\ldots,N$, and $\boldsymbol{\tau}$ (p x 1) is the vector of intercepts. The $\boldsymbol{\eta}_i$ (M x 1) are the latent variables, and $\boldsymbol{\varepsilon}_i$ (p x 1) is the vector of error terms. These variables are independent and have the following distributions $\boldsymbol{\eta}_i \sim N_M(\boldsymbol{\alpha}, \boldsymbol{\Psi})$ and $\boldsymbol{\varepsilon}_i \sim N_p(\mathbf{0}, \boldsymbol{\Theta})$. Lastly, $\boldsymbol{\Gamma}$ (p x M) is the factor loading matrix. Let γ_{jk} be the element in the j^{th} row and k^{th} column of the matrix $\boldsymbol{\Gamma}$; where γ_{jk} is the factor loading between latent variable k ($k = 1,\ldots, M$) and item j ($j = 1,\ldots,p$). The factor loadings (in matrix $\boldsymbol{\Gamma}$) are the regression coefficients for the relationship between observed variables and latent variables. An observed variable is said to load onto a latent variable if the factor loading is non-zero. Researchers can determine which variables load onto which latent variables by factor analysis[38] (EFA or SEM discussed later) or from other research in existing literature.

A type of latent variable model that has been applied to test for various types of RS is confirmatory factor analysis (CFA). CFA is a latent variable modeling technique used to test the relationship between a set of manifest (observed dependent) variables and a set of continuous latent variables. The continuous latent variables are often referred to as factors[38]. CFA includes models where the relationship between latent variables and a set of covariates are studied to understand measurement invariance.

RS is a type of measurement bias; measurement bias is defined as a violation of measurement invariance[39]. Measurement invariance is the notion of consistent measurements across groups or time. There has been plenty of research to determine if measurement invariance holds[37, 56, 59] but there has been little study on how measurement invariance (or lack thereof) will affect factor score estimates.

Factor scores are estimated from latent variable models to provide information about an individual's placement on the latent variable(s)[13]. Since the number of latent variables is, in most cases, smaller than the number of observed variables, it becomes desirable to use the estimated factor scores in subsequent analyses. In theory, by using the estimated factor scores for further analysis, the data dimension is reduced, which will facilitate data handling and modeling[13]. There are various factor scoring methods to estimate factor scores these include: the Thurstone[53] method, Bartlett[4] method, Hoshino and Bentler[25] method, Skrondal and Laake[49] method, and the sum scores method. It is unclear how longitudinal factor scoring estimation performs when measurement non-invariance is unaccounted for in CFA models. For example, if a given longitudinal CFA model is misspecified (non-invariant items are constrained to equality over time), the variance of a latent construct may be biased over time. RS is a type of bias and must be accounted for to avoid contamination of measurement however; the substantive effect of

bias or RS can be meaningful[39]. There is little study on the effects of RS on factor score estimation. Therefore, it is uncertain if RS should be accounted for in the model before estimating factor scores from longitudinal data.

Oort[42] proposed a method to detect RS in a set of data that involves the comparison of two latent variable models; a constrained model and an unconstrained model. A constrained latent model enforces equality in parameters (i.e. factor loadings, intercepts, and residual covariances) across occasion or groups; whereas an unconstrained model has no constraints and parameters are freely estimated. Oort's method is also used to test hypotheses about specific types of RS occurring over time. RS types include: reconceptualization, reprioritization, or recalibration.

According to Oort[42], reconceptualization is present in the data if the pattern of zero and non-zero coefficients in the factor loading matrix differs across time. Then one or more concepts (i.e the latent variables) are defined by unequal sets of observed variables and redefinition of the target construct has occurred. Reprioritization occurs if the value of the factor loading of a particular observed variable changes from one occasion to the next then, that variable has become more or less indicative of the concept involved. Therefore if the factor loading matrix at the first time point differs from the second time point and there is a change in priority of constructs then reprioritization RS has occurred. Recalibration is defined as a change in the respondent's internal standards of measurement. If there is a difference across occasions between intercepts (uniform) or between residual variances (non-uniform) then recalibration RS has occurred.

1.2 Purpose and Objectives

Purpose

For this study the PROs data is continuous for two measurement occasions. Let $\boldsymbol{\alpha}_1$ be the true mean of the latent variables at time one and $\boldsymbol{\alpha}_2$ be the true mean at time two. The null hypothesis of no true change over time is:

$$H_0: \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2.$$

Since the latent variable means cannot be measured, the factor score means can be used to test for true change. Let $\boldsymbol{\mu}_1^f$ be the population mean of the factor scores at time one and $\boldsymbol{\mu}_2^f$ be the population mean at time two. The null hypothesis becomes:

$$H_0: \boldsymbol{\mu}_1^f = \boldsymbol{\mu}_2^f.$$

The purpose of this study is to determine if factor scores can detect true change (i.e. change in the latent variable means) when reprioritization RS exists. Another purpose is to recommend a factor scoring method and CFA model to estimate factor scores when reprioritization RS is present.

Objectives

a) Compare sum scores, Thurstone[53], and Bartlett[4] factor scoring methods performance to detect true change over time with reprioritization RS present via simulation studies by measuring
    i. Type I error
    ii. Statistical power
b) Compare the accuracy of the factor score estimates from CFA models that adjust for reprioritization RS and models that do not by measuring
    i. Factor loading bias

ii. Factor score bias

1.3 Summary

In this study factor scores are used to test for true change in PRO (i.e. change in the latent variable means) when reprioritization RS is present. There are two main problems examined in this study. One, there are various factor scoring methods and it is unknown which factor scores would best detect true change when RS is present. Two, it is unclear how reprioritization RS will affect factor score estimation. Is it best to estimate factor scores from a latent variable model with invariant parameters across time or non-invariant parameters? A simulation study was conducted to determine the optimal latent variable model and factor scoring method to estimate factor scores used to test for true change when reprioritization RS was present. The most optimal model will produce the most accurate factor score estimates. This will be determined by measuring the factor score and factor loading biases. The most optimal factor scoring method will create factor scores that best detect true change. This will be determined by measuring the Type I error rate and statistical power.

The second chapter of this thesis will review definitions and topics related to analysis of longitudinal PROs, factor score estimation, and RS. Chapter 3 will define notation and terms used throughout the study. Also, the simulation methods and parameters will be described. Chapter 4 is a presentation of the results. Lastly, Chapter 5 discusses the implications of the results and simulation study, strengths and limitations of this study, and future work.

Chapter 2: Literature Review

The following topics are reviewed in this chapter: analysis of longitudinal PROs, factor scoring methods, measurement invariance, and RS.

2.1 Analysis of Longitudinal Patient-Reported Outcomes

The study of PROs is a rapidly evolving field of research where the respondent is a patient or person whose experiences (self-report) researchers are interested in. PROs can include information about general health, physical functioning, physical symptoms and toxicity, emotional functioning, cognitive functioning, role functioning, social well-being and functioning, sexual functioning, and existential issues[44]. PROs are answers of patients or persons to questions, referred to as item responses, often grouped into several dimensions, or domains, in a questionnaire. The questionnaires can focus on a single domain, such as physical functioning, or several domains. The sum or average of item responses is known as a summary score and can be used for PRO analysis.

In contrast to observed data, the analysis of PRO data should take into account the latent characteristic of what PROs are intended to measure[5]. However, doing analyses on each domain separately or on just the summary scores can lead to incorrect conclusions. This is caused by ignoring some important information such as: correlations between domains or scores and differences among individual item responses. The most common analysis techniques for PROs include Classical Test Theory (CTT)[5] and latent variable models[42].

The longitudinal CTT method investigates whether a latent outcome changes over time. The method consists of calculating a score by averaging the item responses for each patient and then a linear mixed model is used to explain the evolution of the score with time[5]. The score is a

linear function of the latent outcome and therefore explains evolution of the latent outcome over time.

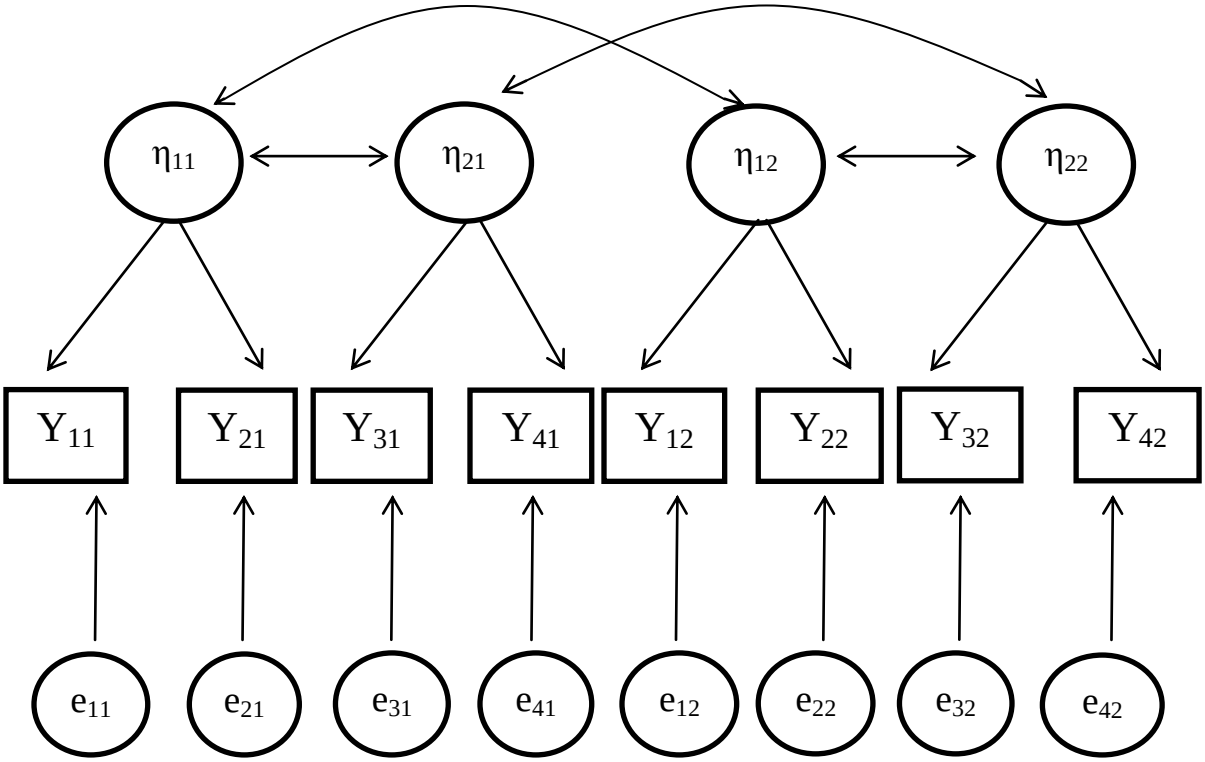


Figure 1: Latent Variable Measurement Model: Graphical display of a latent variable measurement model for individual i; circles represent latent variables and rectangles represent observed variables Y_{jt} for item j at time t. Common factors η_{kt} for factor k (k=1,2) at time t (t=1,2), and residual factors e_{jt} for item j at time t. The single head arrows denote loadings and the double-headed arrows denote correlations.

Another method to analyze PRO data is latent variable models such as structural equation modeling (SEM). SEM is a method used to investigate the relationship between observed variables and a smaller number of latent variables (i.e. the $\boldsymbol{\eta}_i$'s in Figure 1). SEM consists of two parts: a measurement model and a structural model [38]. The measurement model specifies how latent variables are assessed in terms of observed variables (see equation 1.1). The structural model specifies the causal relationship among latent variables[17]. SEM incorporates various analyses such as factor analysis and path analysis as special cases[38].

Factor analysis includes exploratory factor analysis (EFA) and CFA (described in detail in section 1.1). EFA explores possible underlying latent variable structure from a set of observed variables[46] and summarizes the underlying correlation structure for the dataset. EFA is used to identify the underlying factor structure whereas; CFA is used to evaluate hypotheses regarding a specific factor structure. For example, EFA is used to find a model that best fits a dataset. CFA is used to test hypotheses about a dataset for a specific model. Hypotheses could include number of latent variables, relationship of latent variables to observed variables, and correlation between latent variables.

2.2 Factor Scores

Latent variables are defined as random variables whose realizations are not observable. However, it can be advantageous to compute scores of the latent variables for all individuals; these are known as factor scores. Factor scores are estimates of latent variable ($\boldsymbol{\eta}_i$) scores for each individual in a dataset. Let the set of factor scores for the i^{th} individual ($i = 1,\ldots, N$) be denoted as $\mathbf{f}_i$. Consider equation 1.1, the dimension of the latent variables ($\boldsymbol{\eta}_i$) is M x 1 for the i^{th} individual. Similarly, from this equation, the dimension of $\mathbf{f}_i$ will be M x 1 for the i^{th} individual. Figure 1 displays an example of a measurement model for the i^{th} individual. At time 1 ($t = 1$) this model has 2 ($j = 1,2$) latent variables (η_{i11} and η_{i21}) therefore each individual i, at time 1, will have 2 factor scores (fi_{11} and fi_{21}), one for each latent variable.

There are two categories of methods to calculate factor scores; refined and non-refined methods[13]. Non-refined methods are simple summary procedures that are easy to execute and the factor scores are simple to interpret[13]. Refined methods create factor scores using more

sophisticated, complex, and technical approaches[13]. However, with advances in technology the computational complexity of the refined methods is becoming less problematic.

Non-refined factor score methods include partial least squares (PLS) and sum scores. Refined factor score methods include Thurstone[53], Bartlett[4], Skrondal and Laake[49], and Hoshino and Bentler[25]. The PLS, Skrondal and Laake[49], and Hoshino and Bentler[25] methods were not included in this thesis. These methods use factor scores to estimate the relationship between independent and dependent latent variables, which is not the focus of this thesis. Since this thesis focuses only one set of continuous variables the factor scores estimated from the Skrondal and Laake[49] and Hoshino and Bentler[25] method would be the same as the Bartlett method.

2.2.1 Sum Scores Methods

One of the simplest ways to estimate factor scores is the sum scores method. Recall from section 1.1 that factor loadings are the regression coefficients for the observed variable ($\mathbf{Y}_i$) onto the latent variable ($\boldsymbol{\eta}_i$). The factor loadings are the elements of the factor loading matrix, $\boldsymbol{\Gamma}$ (refer to equation 1.1). An observed variable is said to load onto a latent variable if the factor loading (between the observed and latent variable) is non-zero. The sum scores method estimates a factor score by a weighted sum of the observed variables that load onto the corresponding latent variable[9]. If the variable loading onto the latent factor yields a negative factor loading, the contribution of the observed variable is subtracted (rather than added) because the variable is negatively related to the latent variable[13]. This method is simple to use, easy to compute and interpret[13].

In general the observed variables are known as raw scores. If the observed scores are standardized (i.e. observed variables scaled to the sample mean and sample standard deviation)

then they are called standardized scores. Sum scores are calculated by summing the scaled scores of all observed variables loading onto each factor separately. The scaled scores are calculated by multiplying the raw scores by weights (where the weights are between zero and one and sum to one). Choosing the same weights for all raw scores is popular. More details on the calculation of the sum scores is in Chapter 3. Various versions of the sum scores method include only summing raw scores that load onto a factor if the factor loadings are above a cut-off value. Standardized scores could be used instead of raw scores. This is advantageous if standard deviations of the variables exhibit high variation[13].

2.2.2 Thurstone[4] and Bartlett[3] Methods

Thurstone[53] and Bartlett[4] are both refined factor scoring methods and the calculations are described in detail in Chapter 3. The performance of refined methods is measured by three properties: validity, univocality, and correlational accuracy[13]. Since factor scores correspond to latent variables it is advantageous for the factor scores ($\mathbf{f}_i$) and the corresponding latent variables ($\boldsymbol{\eta}_i$) to be correlated, this property is known as validity[13, 21]. Also, it is desirable for the factor scores to have the same correlation pattern (in terms of zero and non-zero correlation) as the latent variables. For example, if two latent variables are uncorrelated (i.e. have a correlation of zero) then the corresponding factor scores (for each individual) should also be uncorrelated, this is known as correlational accuracy[13, 21]. Moreover, if two latent variables are uncorrelated then the corresponding factor score should be uncorrelated to the latent variable. For example consider Figure 1, if η_{12} and η_{22} are uncorrelated then the factor score f_{12} (corresponding to η_{12}) for individual i will be uncorrelated to the latent variable η_{22}, this is known as univocality[13, 21]. Ideally a refined factor scoring method should generate factor score estimates that satisfy all

these properties. There has yet to be a factor scoring method able to produce factor score estimates that have validity, univocality, and correlational accuracy properties. The Bartlett[4] and Thurstone[53] methods are more sophisticated and computation is more complex than the sum scores method. Unlike the sum scores method, these methods retain the correlation relationships between factors.

Thurstone Method:

Thurstone's[53] method has been the most commonly utilized refined method[29]. The Thurstone method[53] uses a least squares regression approach to predict factor scores[13]. Let factor scores estimated from the Thurstone method[53] be known as Thurstone factor scores. The primary advantage of this method is that the Thurstone factor scores maximize validity[13]. Thurstone[53] factor scores are not unbiased estimates of true factor scores[25] (see appendix). Also the Thurstone factor scores are not univocal[13].

Hoshino and Bentler[25] noted Thurstone[53] factor scores always have variances less than unity[29] which generally leads to correlational inaccuracy (i.e. correlational accuracy does not hold). It was once thought that Thurstone factor scores would be the most suitable as independent variables in regression analysis[29] (described in detail in section 2.2.3). However, recent simulation studies, such as Lastovicka's[29], have shown that using Thurstone factor scores as independent variables in regression analysis often results in biased estimates of the regression coefficients. Lastovicka[29] concluded that under less desirable data conditions, such as incorrect assumptions of the latent variable correlations, high uniqueness (i.e. the variance of most of the observed variables are very high), and low number of observed variables, the Thurstone factor scores should not be used as independent variables in regression models[29].

Bartlett Method:

The Bartlett[4] method is a maximum likelihood method obtained by minimizing the sum of squares of the latent variables across all observed variables. Let factor scores estimated from the Bartlett[4] method be known as Bartlett factor scores. Statistical derivations have shown that this method produces unbiased estimates of the true factor scores[24]. However, the Bartlett[4] factor scores can result in biased regression coefficient estimates. Similar to the Thurstone factor scores the Bartlett factor scores also have high validity[20]. However, unlike the Thurstone method, this method has the additional advantage that the factor scores are univocal[13]. Also, the Bartlett method assumes that the latent variables are uncorrelated. This can result in less accurate estimates of factor scores when latent variables are highly correlated. Lastly, a disadvantage of the Bartlett factor scores is correlational inaccuracy.

2.2.3 Applications of Factor Scores

It is common to use factor scores, as opposed to the original data, in subsequent analyses such as regression analysis, predictive analysis, and cluster analysis[7, 13, 19, 58]. Factor scores are popular because the number of latent variables is usually smaller than the number of observed variables. By using factor scores instead of observed variables the data dimension is reduced. Also, the factor scores provide numeric values to the un-measureable latent constructs (i.e. mental health).

A popular use of factor scores is the factor score regression (FSR) method[25]. The FSR method tests the association between latent variables using the estimated factor scores. In FSR the factor scores are estimated for the independent and dependent observed variables and used in regression analysis. In fact, FSR has been regarded as a major application of factor score

estimates[49]. Hoshino and Bentler[25] note that the factor scoring method (i.e. sum scores, Bartlett[4], Thurstone[53], etc.) used to calculate the factor scores could result in biased regression coefficients (refer to the appendix).

Factor scores have been also used in a variety of analyses. Boomsma[7] used factor analysis to decompose traits of twins into genetic and environmental factor scores. Knowledge of these environmental and genetic factor scores could help the health community better understand disease risks (such as blood pressure)[7]. Wu et al.[58] performed cluster analysis on factor scores to infer how family interactions are associated with depressive symptoms in children. Goldstein's[19] paper applied factor analysis to data from a survey of homeless veterans and factor scores were estimated. These factor scores were then used to infer the association between obtaining a large factor loading on any latent variable (cardiac, mood, stress, addiction and psychosis) with a number of sociodemographic and homelessness related variables. If the factor scores are estimated from longitudinal data, then change in factor scores over time can be tested by longitudinal analyses[7].

Simulation studies[25, 29, 49] have been conducted to compare differences between factor scoring methods, in terms of finite sample performance, estimation bias and standard errors, and multiple R (the correlation coefficient between the observed and predicted values). Hoshino and Bentler[25] and Skrondal and Laake[49] conducted simulation studies to compare their methods with other factor scoring methods used in FSR. The bias of the regression coefficient estimates between independent and dependent latent variables was of interest in these studies. Skrondal and Laake[49] proved mathematically that the Thurstone[53] and Bartlett[4] method will produce biased regression coefficient estimates (see appendix). Lastovicka[29] tested many methods under a variety of data conditions (latent variable variance, latent variable correlation assumptions, and

number of observed variables) for the case of using factors as independent variables in regression. Researchers have used factor scores as independent variables in regression analysis because of data reduction and simplification, to avoid problems such as multicolinearity, and determine relationships between latent variables and observed variables. Lastovicka[29] concluded that depending on how the factor scores are to be used in subsequent analyses dictates which method is best.

2.3 Measurement Invariance

Measurement invariance is crucial because comparisons and analyses (across groups or occasions) are meaningful only if the same construct is being measured[56]. Measurement invariance is a property of measurement that the same construct is being measured across groups or occasions[59]. Widaman et al. explains "Investigating whether the same construct is assessed on the same metric across groups or occasions is under the rubric of measurement invariance."[56]. For example, consider a latent variable model; measurement invariance implies that the loadings for the observed variables on a single underlying factor across groups or occasions are on the same metric and stronger conclusions are warranted[56]. Although extensive work has been published on this topic, the majority of this work has focused on invariance across groups, with less emphasis on exploring invariance over time. However, model constraints to investigate measurement invariance across occasions are the same as across groups, the difference being that constraints are applied over time rather than between groups. Also, data across occasions can be viewed as several group samples that are dependent.

There are four types of measurement invariance across occasions which are:

a) Configural invariance: the same pattern (i.e. zero and non-zero values) of fixed and free factor loadings across time.
b) Weak invariance: factor loadings remain constant across time.
c) Strong invariance: factor loadings and intercepts remain constant across time.
d) Strict invariance: factor loadings, intercepts, and residual variances (i.e. diagonal entries of $\boldsymbol{\Theta}$ in equation 1.1) remain constant across time[59].

Measurement invariance is tested for two nested measurement models. Two models are nested if both have the same parameters and one model has at least one additional parameter. The χ^2 statistic is calculated from the difference in the log of the likelihood function between the two nested models. The test statistic asymptotically follows a χ^2 distribution, with the degrees of freedom (df) equal to the difference in df of the two models. A non-significant difference in fit is considered as evidence for measurement invariance[59]. Model fit is compared sequentially for:

a) configural invariance model vs. weak invariance model
b) weak invariance model vs. strong invariance model
c) strong invariance model vs. strict invariance model.

If there is a non-significant difference in fit then the next pair of models is compared until there is a significant difference in fit or until strict invariance is established.

To claim the same latent construct is measured over time, strong or strict invariance must hold across time[56] but strict invariance is preferred[37]. This is because maintaining strict invariance establishes that any differences observed over time are the sole function of the means, variances, and covariances of the construct over time[37].

Measurement bias is formally defined as a violation of measurement invariance. RS can be understood as a one type of measurement bias, where we want to measure change in the latent construct but changes in the construct are not fully captured by changes in the observed variables[43].

2.4 Response Shift

RS is a potential source of bias in longitudinal PRO studies. When RS is present in data, conventional statistical analysis methods may not detect true change in PROs, even when a true change exists in the population[31]. RS may not affect all PROs or patient groups equally and can result in low statistical power and biased conclusions about true change in a population. The definition of RS coincides with the general definition of measurement invariance, violation across occasions[58]. There are an increasing number of studies on how to identify the presence of RS in PROs[16].

There are many methods to detect RS[31, 35, 42]. One method is Oort's SEM method[42]. Oort[42] developed a SEM method which can not only detect various types of RS, but also measure true change[42]. Oort's procedure consists of four steps; it has been developed for measuring RS in data collected at two time points[42]:

1) Establish a latent variable model that has good model fit (statistics such as root mean square error approximation (RMSEA), χ^2, comparative fit index **(**CFI), standardized root mean square residual (SRMR)) and has a clear interpretation called Model 1. This model has identification constraints of latent variable means and variances equal to zero and one respectively. If the patterns of zero and non-zero factor loadings (**Γ**) across occasions in Model 1 are very different (i.e. none of the factor loading patterns are the same across

occasion), then reconceptualization RS has occurred and no further testing is conducted. Otherwise if the model has good fit and the patterns of the factor loadings are largely the same then the procedure continues to step two.

2) Fit a "no RS model" also known as a constrained model and call this Model 2. This model imposes constraints over invariance hypotheses associated with RS (i.e. intercepts, factor loadings, and residual covariances). If the fit of Model 2 is not significantly worse than Model 1, then all changes in the observed means and covariances can be explained by the latent variable means and covariances. The χ^2 test can be used to test for no RS. If there is no significant difference in fit then there is no RS and Step 3 can be skipped. However, if there is a significant difference in fit then RS is present in the data. Step 3 will detect the type of RS present so true change can be measured.
3) From the previous step we have determined whether there is RS in the data. This step includes a step by step removal of all untenable constraints and yields a model in which all apparent RS are accounted for in the factor loadings, intercepts, and residual covariances. This is known as Model 3 and is guided by expected parameter changes, Wald tests, inspection of the standardized discrepancies between the estimated and observed variable ($\mathbf{Y}_i$ from equation 1.1) means, variances, and covariances, modification indices (MOD). MOD reflects the improvement in model fit if a specific coefficient is estimated rather than fixed. Subsequently, the fit improvement of each modification step can be tested through the χ^2 test. However, the model should not be modified in ways that make no sense (i.e. variables at time 2 should not load on latent variables from time 1).
4) From Step 3 we have a final model (Model 3) and from this model true change can be measured. The χ^2 statistic is used to test for true change. If the null hypothesis of

invariant latent variables means (i.e. the latent variable means are equal) is rejected, then the difference between the latent variable means across occasion can be taken as a measure of true change. Otherwise, there is no change over time between the latent variable means. Other changes, such as true change in the variances and correlations of the latent variables, can be tested by adding across occasion constraints to the latent variable covariances and correlations or residual correlations.

2.5 Summary

Methods to analyze PROs include CTT and various factor analytic methods; including SEM. Ignoring RS has the potential to lead to bias in the resultant analysis. RS is a violation of measurement invariance across occasions. There have been many methods developed to detect RS in longitudinal PROs[31, 35, 42] but little on the effects of subsequent analyses when RS is identified. Factor scores provide information about an individual's placement on latent variables. Factor scoring methods are numerous[1, 4, 12, 13, 25, 29, 30, 49, 53] and have been compared in cross sectional studies[25, 29, 49], but not in studies over time. Factor score estimation and inferences about factor scores could be affected by the presence of RS in the data. It is unclear how constraining non-invariant parameters to be equal could affect the estimation of factor scores.

Chapter 3: Methods

This chapter introduces notation and models that are used to define the factor scoring methods and describe the simulation study.

3.1 Notation

Let Y_{ijt} represent a dependent variable for the i^{th} study participant (i=1, …, N), for the j^{th} item (j=1, …, p) at time t (t=1, …, T). The mean and variance of Y_{ijt} are represented by $E(Y_{ijt}) = \mu_{ijt}$ and $Var(Y_{ijt}) = v_{ijt}$. For this study, let the number of items $p = 4$ and the number of latent variables $M = 1$. Let the measurement model, for each individual i, be defined for $t = 2$ time points:

$$\mathbf{Y}_i = \mathbf{\Gamma}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_{i,} \tag{3.2}$$

where $\mathbf{Y}_i = \begin{pmatrix} \mathbf{Y}_{i1} \\ \mathbf{Y}_{i2} \end{pmatrix}$ is the 8 x 1 vector of observed variables (with $\mathbf{Y}_{i1}$ and $\mathbf{Y}_{i2}$ are both 4 x 1) with covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} \\ \mathbf{\Sigma_{12}^T} & \mathbf{\Sigma_{22}} \end{pmatrix}$ (of dimension 8 x 8) where $\mathbf{\Sigma_{11}} = \text{cov}(\mathbf{Y}_{i1}, \mathbf{Y}_{i1})$, $\mathbf{\Sigma_{12}} = \text{cov}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2})$, and $\mathbf{\Sigma_{22}} = \text{cov}(\mathbf{Y}_{i2}, \mathbf{Y}_{i2})$ and all these matrices have dimension 4 x 4.. The matrix of regression coefficients of $\mathbf{Y}_i$ on $\boldsymbol{\eta}_i$ is $\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma_2} \end{pmatrix}$ (of dimension 8 x 2) known as the factor loading matrix, where $\mathbf{0}$, $\mathbf{\Gamma_2}$, and $\mathbf{\Gamma_1}$ are all of dimension 4 x 1. Let the elements of the matrix $\mathbf{\Gamma}_t$ (t = 1, 2) be denoted as γ_{jmt} for $1 \leq j \leq p$ and $1 \leq m \leq M$. The $\boldsymbol{\eta}_i = \begin{pmatrix} \eta_{i1} \\ \eta_{i2} \end{pmatrix}$ is the 2 x 1 random vector of latent variables with mean $\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$ (2x1) and covariance matrix $\mathbf{\Psi} = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{12}^T & \Psi_{22} \end{pmatrix}$ (2 x 2) where $\Psi_{11} = \text{cov}(\eta_{i1}, \eta_{i1})$, $\Psi_{12} = \text{cov}(\eta_{i1}, \eta_{i2})$, and $\Psi_{22} = \text{cov}(\eta_{i2}, \eta_{i2})$. The measurement error vector is

$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \boldsymbol{\varepsilon}_{i1} \\ \boldsymbol{\varepsilon}_{i2} \end{pmatrix}$ of dimension 8 x 1 with mean $\mathbf{0}$ (8 x 1) and covariance matrix $\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\Theta}_{12} \\ \boldsymbol{\Theta}_{12}^{T} & \boldsymbol{\Theta}_{22} \end{pmatrix}$ (8 x 8)

where $\boldsymbol{\Theta}_{11}$= cov($\boldsymbol{\varepsilon}_{i1}$, $\boldsymbol{\varepsilon}_{i1}$), $\boldsymbol{\Theta}_{12}$= cov($\boldsymbol{\varepsilon}_{i1}$, $\boldsymbol{\varepsilon}_{i2}$), and $\boldsymbol{\Theta}_{22}$= cov($\boldsymbol{\varepsilon}_{i2}$, $\boldsymbol{\varepsilon}_{i2}$) and all these matrices have dimension 4 x 4. It is assumed that $\boldsymbol{\varepsilon}_i$ is uncorrelated with $\boldsymbol{\eta}_i$. In this study the vector of intercepts $\boldsymbol{\tau}$ are set to zero which is why $\boldsymbol{\tau}$ has been dropped from equation 3.2.

Consider the model defined by equation 3.2. From this model the covariance matrix of $\mathbf{Y}_i$ is estimated as follows:

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Psi}\boldsymbol{\Gamma}^{T} + \boldsymbol{\Theta}, \tag{3.3}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the dependent observed variables, $\boldsymbol{\Psi}$ is the covariance matrix of the unobserved common latent variables, $\boldsymbol{\Theta}$ is the covariance matrix of the residuals.

3.2 Factor Scoring Methods for the Simulation Study

The factor scoring methods investigated in the simulation study included: sum scores method, Thurstone[53] method, and Bartlett[4] method.

3.2.1 Sum Scores Methods

In this study only CFA models are used to analyze the data. As mentioned in section 2.1 CFA is used to evaluate hypotheses regarding a specific factor structure. Therefore the factor structure is determined beforehand (through various methods mentioned in section 1.1) and it is known which observed variables load onto each latent variable.

Let γ_{jmt} be the population regression coefficient between the j^{th} item and m^{th} latent variable at time t. Where $\gamma_{jmt} = 0$ implies that Y_{ijt} does not load onto latent variable m for each individual i= 1, …, N. Otherwise, if $\gamma_{jmt} \neq 0$ then Y_{ijt} loads onto latent variable m. The sum scores factor score for individual i and latent variable m at time t is estimated as follows:

$$f_{imt} = \sum_{j=1}^{p} I(\gamma_{jmt} \neq 0) w_j \kappa_j Y_{ijt}, \quad (3.4)$$

where $I(\gamma_{jmt} \neq 0) = 1$ when $\gamma_{jmt} \neq 0$ and 0 otherwise and w_{jmt} ($0 < w_{jmt} < 1$) are constants such that $\sum_{j=1}^{p} w_{jmt} = 1$. Also, $\kappa_{jmt} = -1$ if $\gamma_{jmt} < 0$ and $\kappa_{jmt} = 1$ if $\gamma_{jmt} > 0$.

3.2.2 Bartlett and Thurstone Methods

The Bartlett and Thurstone factor scores are estimated by[3]

$$\mathbf{f}_i = \mathbf{A}^T \mathbf{Y}_{it}, \quad (3.5)$$

where $\mathbf{f}_i$ is the vector of factor scores for the i^{th} individual with dimension m x t, and $\mathbf{A}$ is a matrix of weights that are constant across subjects.

Thurstone Method[53]:

For theThurstone[53] method one derives the weight matrix $\mathbf{A}$ for uncorrelated latent variables as:

$$\mathbf{A} = \mathbf{\Theta}^{-1}\mathbf{\Gamma}(\mathbf{I} + \mathbf{\Gamma}^T\mathbf{\Theta}^{-1}\mathbf{\Gamma}\mathbf{\Psi})^{-1}\mathbf{\Psi}, \quad (3.6)$$

or if the latent variables are correlated,

$$\mathbf{A} = \mathbf{\Theta}^{-1}\mathbf{\Gamma}\,(\mathbf{\Psi}^{-1} + \mathbf{\Gamma}^T\mathbf{\Theta}^{-1}\mathbf{\Gamma})^{-1}, \quad (3.7)$$

where, $\mathbf{I}$ is the identity matrix. The researcher must determine which weighting matrix to use (i.e. if the latent variables are correlated or not). This can be determined by existing literature, additional analysis (such as factor analysis), or common knowledge. This research examines RS and therefore looks at factor scores over time. It is reasonable to assume that latent variables are correlated over time therefore the weighting matrix from equation 3.7 will be used.

Bartlett Method[4]:

The Bartlett[4] weight matrix is derived as

$$\mathbf{A} = \mathbf{\Theta}^{-1}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Theta}^{-1}\mathbf{\Gamma})^{-1}. \quad (3.8)$$

This weight is derived under the assumption that the latent factors are not correlated and could result in less accurate estimates of factor score if there is a strong correlation between latent variables[7].

A summary of the factor scoring methods can be found in Table 1 and Table A1. Specifically, Table 1 summarizes the methods and describes their advantages and disadvantages, while Table A1 (appendix) provides statistical derivations for the Thurstone[53] and Bartlett[4] factor scoring methods.

3.3 Simulation Studies

3.3.1 Simulation Design

A series of Monte Carlo simulations were conducted to evaluate the effects of RS on estimated factor scores. All simulation studies were done in the following steps:

1) Generate multivariate data for each simulation scenario described in section 3.3.3;
2) Analyze the generated data with three different CFA models with different parameter constraints invariant over time. Details are described in section 3.3.2;
3) Calculate factor score estimates using the described factor scoring methods. Test change of factor score mean with a paired t-test for each factor scoring method;
4) Evaluate the performance of factor scoring methods in terms of bias in parameters of interest (i.e. factor loading and factor score mean), Type I error rate, and statistical power.

All simulations were done in SAS and SAS/IML version 9.3 using PROC CALIS for the CFA model analysis. In this study, the simulated data was multivariate and generated based on

equation 3.2 with known latent variable mean ($\boldsymbol{\alpha}$) and covariance ($\boldsymbol{\Psi}$) and residual covariance ($\boldsymbol{\Theta}$). The latent variables ($\boldsymbol{\eta}_i$'s) and error terms ($\boldsymbol{\varepsilon}_i$'s) were generated from multivariate normal distributions and were used to generate the observed variables ($\mathbf{Y}_i$). The error terms and the latent variables were generated as $\boldsymbol{\varepsilon}_i \sim N_8(\mathbf{0}, \boldsymbol{\Theta})$ and $\boldsymbol{\eta}_i \sim N_2(\boldsymbol{\alpha}, \boldsymbol{\Psi})$. All variables are defined in section 3.3.3.

Table 1: Advantages and Considerations of Common Factor Scoring Methods

Name	Description	Advantages	Considerations
Weighted Sum Scores	Take into consideration the loading values in the factor score and multiply the factor loading to the scale score then sum (cutoff can be applied)[13]	Recognizes the strength of items and items with highest loadings have the most effect on the factor scores[13]	Possibility that differences in factor loadings are due to extraction and rotation choices and in that case this method is not better than sum scores[25]
Thurstone	Multiple regression used to estimate factor scores (available in SAS, SPSS, and R)[13]	Factor scores are standard scores with mean zero and variance SMC between items and factors and maximizes validity of estimates[13]	Factor scores are neither univocal nor unbiased and scores may be correlated even when factors are orthogonal[27]
Bartlett	Method of producing factor scores most likely to represent the true scores (available in SPSS or R)[13]	Factor scores are standard scores, the estimates are unbiased and univocal, and the procedure produces high validity between scores and factor[7, 13]	The factor scores may be correlated even when factors are orthogonal[7,13]

SD = standard deviation, SMC = squared multiple correlation, standard scores = mean =0, variance = SMC.

This study examines the effect of RS on factor scores when factor scores are used to test equality of latent variable mean ($\boldsymbol{\alpha}$) across time (i.e. $\alpha_1 = \alpha_2$). For measuring Type I error the population latent variable mean of the generated data were equal ($\alpha_1 = \alpha_2$) and for the power rate the means were not equal ($\alpha_1 \neq \alpha_2$). Two different latent variable means were considered $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_{05}$ where $\boldsymbol{\alpha}_0 = \begin{pmatrix} 0 \\ \alpha_2 \end{pmatrix}$ and $\boldsymbol{\alpha}_{05} = \begin{pmatrix} 0.5 \\ \alpha_2 \end{pmatrix}$. The value of α_2 depends on whether Type I error or power rate was being measured. The values of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_{05}$ are defined in more detail in section 3.3.3.

In this study none of the factor loadings were constrained to be 1. Therefore, for model identification, the variances of the latent variables ($\Psi_{11}$and Ψ_{22}) were constrained to be 1. Since this simulation study has only one latent variable ($m = 1$) at both time points then the subscript m will be dropped from previous notations (i.e. f_{imt} becomes f_{it} and γ_{jmt} becomes γ_{jt}) in section 3.3.

3.3.2 Simulation Study Definitions and Terms

The simulations modeled unconstrained, constrained, and partially constrained CFA models. The unconstrained model had no constraints on factor loadings, intercepts, or residual variances across time. This model is the same as the configural invariance model described in chapter 2. The partially constrained model held the factor loadings ($\boldsymbol{\Gamma}$) equal across time; this is the same as the weak invariance model. Lastly, the fully constrained model held the factor loadings ($\boldsymbol{\Gamma}$) and the residual variances ($\text{Var}(\varepsilon_{ijt})$) equal across time; same as the strict invariance model.

The influence of RS on factor scoring estimation is investigated in terms of Type I error rate, power, and bias calculations of factor score means and factor loadings. The factor scoring method least affected by RS in relation to these measures will be considered the optimal factor scoring method in the presence of RS. Let the constrained Bartlett and Thurstone factor scores

denote factor scores calculated from the parameter estimates of the constrained CFA model. Let the unconstrained Bartlett and Thurstone factor scores denote factor scores calculated from the unconstrained CFA model estimates. Let the partially constrained Bartlett and Thurstone factor scores denote factor scores calculated from the partially constrained CFA model estimates. Let the constrained sum scores factor scores be factor scores calculated such that the weights are equal across time. The constrained sum scores factor scores for this study was calculated as follows:

$$f^{c}_{it} = (\sum_{j=1}^{4} \kappa_{jt} Y_{ijt})/4, \quad (3.9)$$

where the superscript c denotes the constraint. Let the unconstrained sum scores method be factor scores calculated such that the weights are not equal across time. The unconstrained factor scores were calculated as follows:

$$f^{u}_{i1} = (\kappa_{11} Y_{i11} + \kappa_{31} Y_{i31})/2 \quad (3.10)$$

$$f^{u}_{i2} = (\kappa_{22} Y_{i22} + \kappa_{42} Y_{i42})/2, \quad (3.11)$$

where the superscript u denotes the lack of constraint. There was no partially constrained sum scores method since the sum scores were estimated from the observed data and not from a latent variable model.

The null hypothesis is that the latent variable mean is equal across time. This hypothesis is examined by testing equality of factor score means across time. The effect size (ES) is used to measure the degree to which the null hypothesis is false. The ES is some specific positive value in the population. The larger the ES the greater the degree to which the phenomenon under study is manifested. In this study the ES is the difference between the true factor mean at the two time

points. Since this study had only one latent variable at each time the ES is a scalar value. The ES based on Cohen's definition[8] is defined as the absolute difference:

$$d = |\alpha_1 - \alpha_2|, \tag{3.12}$$

where d is the ES, α_1 is the latent variable mean at time 1, and α_2 at time 2. Details of the values of d examined in this study are in section 3.3.3.

The RS magnitude was calculated as the norm difference between the factor loadings at time point 1 and 2.

$$S = \|\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2\|, \tag{3.13}$$

Where S is the RS, $\mathbf{\Gamma}_1$ denotes the factor loading vector at time one, and $\mathbf{\Gamma}_2$ denotes the factor loading vector at time two. Details on the values of S examined in this study are in section 3.3.3.

For each factor scoring method factor scores f_{it} were calculated for each individual $i = 1,\ldots,N$ at time point t= 1, 2. Let $\mu_1^f = E(f_{i1})$ (i.e. the expected value of the population factor score at time 1) and $\mu_2^f = E(f_{i2})$. Let $\bar{f}_t$ denote the sample factor score mean for the latent variable at time t. Recall the null hypothesis, that there is no change over time, from section 1.2.

$$H_o: \mu_1^f = \mu_2^f.$$

This hypothesis was tested with the paired-t test. Based on the number of rejections of the null hypothesis Type I error rate (when the null hypothesis is correct) and statistical power (for the case when the null hypothesis is false) were calculated for each scoring method.

The Type I error rate, a, is the rate of rejecting a true null hypothesis. The Type II error rate, b, represents the rate of failing to reject a false null hypothesis. Statistical power is 1-b and is defined as the probability of rejecting the null hypothesis when the null hypothesis is false.[8] The Type I error rate was calculated for d = 0 by counting the proportion of falsely rejecting the

null hypothesis. The power was calculated by counting the proportion of rejections of the null hypothesis for $d > 0$.

Bias is the difference between the expected value of an estimator and the true value of the parameter. The bias for the factor score means and factor loadings were calculated for all simulation replications. Factor loading bias was calculated since factor loadings are affected by reprioritization RS and were used to calculate the refined factor scores. Let $r = 1,\ldots, R$ be the number of simulation replications and superscript r be the sample estimate corresponding to simulation replication r. For the r^{th} generated data set, the relative bias, and average relative bias, for the factor scores and factor loadings were calculated as follows:

$$\text{Rel Bias}_{\text{fl}} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\gamma}_{jt}^{(r)} - \gamma_{jt}\right)/\gamma_{jt}, \tag{3.14}$$

$$\text{Rel Bias}_{\text{fs}} = \frac{1}{R}\sum_{r=1}^{R}\left(\bar{f}_t^{(r)} - \mu_t^f\right)/\mu_t^f, \tag{3.15}$$

where $\hat{\gamma}_{jt}^{(r)}$ represents the sample estimate of the factor loading for simulation r. Average relative bias for the factor loadings were calculated as follows:

$$\overline{\text{Rel Bıasfl}} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{pT}\sum_{t=1}^{T}\left(\sum_{j=1}^{p}\left(\hat{\gamma}_{jt}^{(r)} - \gamma_{jt}\right)/\gamma_{jt}\right)\right) \tag{3.18}$$

3.3.3 Simulation Parameters

The study generated data with reprioritization RS for 1000 replications (R); the simulation parameters were as follows:

a) Sample size: N = 200, 500, 1000

b) Factor loading values:

$\boldsymbol{\Gamma}\text{eq}^{\text{T}}=$

$$\begin{bmatrix} 0.6 & 0.6 & 0.6 & 0.6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.6 + S/\sqrt{8} & 0.6 + S/\sqrt{2} & 0.6 + S/2 & 0.6 + S/\sqrt{8} \end{bmatrix},$$

$\boldsymbol{\Gamma}\text{uneq}^{T}$ =

$$\begin{bmatrix} 0.7 & 0.6 & 0.5 & 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.7 + S/\sqrt{8} & 0.6 + S/\sqrt{2} & 0.5 + S/2 & 0.4 + S/\sqrt{8} \end{bmatrix}$$

c) Mean of latent variables: $\boldsymbol{\alpha}_0 = [0 \quad (0 + d)]^T$ and $\boldsymbol{\alpha}_{05} = [0.5 \quad (0.5 + d)]^T$

d) Factor scoring method: Bartlett, Thurstone, and sum scores

e) S = 0 (none), 0.2 (small), 0.5 (moderate), 0.8 (large)

f) d = 0 (none). 0.05 (medium), 0.1 (large)

g) Model: unconstrained (uncon), fully constrained (con), partially constrained (pcon).

This research focuses on continuous data for two time points since most research for RS[31, 35, 42, 47, 48] examines this type of data. The three measurement models unconstrained, constrained, and partially constrained were chosen because these models are used to detect RS in continuous data[42]. The RS magnitudes (0, 0.2, 0.5, 0.8) were based on Oort's paper[42] and the model had only one latent variable for simplicity. The ES values were chosen from testing simulated data without RS. The values represent large, medium, and no ES. The factor score methods Bartlett, Thurstone, and sum scores methods were selected because these methods are most common and simple to program since they have been included in many simulation studies[25, 29, 49]. The factor loading and sample sizes were chosen from other simulation studies[25, 29, 49]. There were two choices of latent variable means (zero and non-zero). For both zero and non-zero latent variable mean, when estimating the Type I error the latent variable means were constant across time. When estimating the power the latent variable means were not constant. The common assumption is that the latent variable means are zero and it is unknown whether a non-zero mean will affect the factor score estimates.

The common factor and residual covariance matrix, $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$ values were selected based on previous research[25, 29, 49] and resulted in a positive definite covariance matrix, $\boldsymbol{\Sigma}$, see equation 3.3. They are defined as:

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}, \text{ and } \boldsymbol{\Theta} = \begin{bmatrix} 0.5 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & & 0.5 & 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0.2 \\ 0.2 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0.5 \end{bmatrix}$$

Chapter 4: Simulation Results

The performance of the Bartlett, Thurstone, and sum scores methods in the presence of RS was compared by simulation. The comparison is based on Type I error, power, and relative bias (for both factor loadings and the factor score means).

Three types of latent variable models were used to estimate the factor scores: unconstrained (Uncon), partially constrained (Pcon), and fully constrained (Con) models. Therefore there are three types of Thurstone and Bartlett factor scores (Uncon, Con, and Pcon). However there were only two types of sum scores factor scores, the constrained (under Pcon) and unconstrained. Sum scores method is estimated from the observed data and not from estimates of the latent variable model and only two weighting methods were used (fully constrained f^{c}_{it} and unconstrained f^{u}_{it}). These methods were described in detail in section 3.3.2.

The Type I error rates, along with a 95% confidence interval, for the factor scoring methods were calculated for the two latent variable means, three latent model types, three sample sizes, and four RS magnitudes. The power rates for the factor scoring methods were estimated for each model, RS magnitude, latent variable mean, sample size, and two ES values. The relative bias for the factor loadings are calculated for all RS values, both factor loading values, and the unconstrained and constrained latent models. Lastly the relative bias for the factor scores are calculated for the factor scoring method, factor loading values, latent variable means, sample size, RS magnitude, and latent variable models.

4.1 Type I Error Rates

Tables 2 and 3 display the Type I error rates (i.e. the rejection rates for $d = 0$) for the factor scoring methods. The tables display the results for unequal factor loadings ($\boldsymbol{\Gamma}_{uneq}$). The results for equal factor loadings ($\boldsymbol{\Gamma}_{eq}$) were very similar and thus not reported. Across the factor scoring methods the rejection rates, when the latent variable mean was zero ($\boldsymbol{\alpha}_0$), was approximately 0.05 regardless of sample size and RS magnitude. However, for non-zero latent variable mean ($\boldsymbol{\alpha}_{05}$) this was not the case. Type I error rate was influenced by sample size, RS magnitude, and factor scoring method.

Under a constrained model (Con and Pcon) for $\boldsymbol{\alpha}_{05}$ the Type I error rates were similar for all factor scoring methods. When RS was absent ($S = 0$) the Type I error rates were approximately 0.05. However, when RS was present ($S \neq 0$) the Type I error rate increased as sample size and RS magnitude increased. For instance, from Table 3 with a sample size of 200 as RS increased from 0, 0.2, 0.5, to 0.8 the error rates increased from 0.05, 0.22, 0.78, to 0.96 respectively. Similarly for sample size 500: 0.05, 0.52, 0.99, and 1; and for 1000: 0.05, 0.79,1, and 1.

Under the unconstrained model for $\boldsymbol{\alpha}_{05}$, Type I error rates depended on the factor scoring method. The unconstrained Bartlett method had rejection rates between 0.09-0.10 regardless of RS magnitude and sample size. For the Thurstone and the sum scores method Type I error rates increased as RS and sample size increased. Consider Table 3, the error rates of the Thurstone method when sample size was 200 as RS increased from 0, 0.2, 0.5, to 0.8 were: 0.070, 0.085, 0.195, and 0.278 respectively. Similarly, when sample size was 500 as RS increased from 0, 0.2, 0.5, to 0.8 the Type I error rates were: 0.056, 0.138, 0.377, and 0.550. In general the error rates were larger for larger sample size and RS magnitude. The sum scores Type I error rates also

increased as RS magnitude and sample size increased and in general had larger Type I error rates compared to the other factor scoring methods.

The only factor scoring method not affected by RS was the unconstrained Bartlett method. The Type I error rate increased as RS increased for all other methods. Even though true change was zero the hypothesis tests detected change when RS was present and therefore, the significance test results were related to RS, not the true change. Although the unconstrained Bartlett Type I error rates were not affected by RS the rates were still high (0.10) especially when RS was absent.

Table 2: Type I Error Rates for Γ_{uneq} and α_0

		S = 0					
		N = 200		*N* = 500		*N* = 1000	
Model	**Factor Method**	**Rate**	**95% CI**	**Rate**	**95% CI**	**Rate**	**95% CI**
Uncon	Bartlett	0.053	(0.039, 0.067)	0.040	(0.028, 0.052)	0.053	(0.039, 0.067)
	Thurstone	0.052	(0.038, 0.066)	0.040	(0.028, 0.052)	0.050	(0.036, 0.064)
	Sum Scores	0.051	(0.037, 0.065)	0.049	(0.036, 0.062)	0.050	(0.036, 0.064)
Pcon	Bartlett	0.053	(0.039, 0.067)	0.042	(0.030, 0.054)	0.052	(0.038, 0.066)
	Thurstone	0.054	(0.040, 0.068)	0.041	(0.029, 0.053)	0.051	(0.037, 0.065)
	Sum Scores	0.045	(0.032, 0.058)	0.041	(0.029, 0.053)	0.051	(0.037, 0.065)
Con	Bartlett	0.053	(0.039, 0.067)	0.043	(0.030, 0.056)	0.050	(0.036, 0.064)
	Thurstone	0.053	(0.039, 0.067)	0.043	(0.030, 0.056)	0.050	(0.036, 0.064)
				S = 0.2			
Uncon	Bartlett	0.045	(0.032, 0.058)	0.044	(0.031, 0.057)	0.050	(0.036, 0.064)
	Thurstone	0.045	(0.032, 0.058)	0.041	(0.029, 0.053)	0.049	(0.036, 0.062)
	Sum Scores	0.048	(0.035, 0.061)	0.049	(0.036, 0.062)	0.049	(0.036, 0.062)
Pcon	Bartlett	0.044	(0.031, 0.057)	0.037	(0.025, 0.049)	0.053	(0.039, 0.067)
	Thurstone	0.045	(0.032, 0.058)	0.036	(0.024, 0.048)	0.052	(0.038, 0.066)
	Sum Scores	0.041	(0.029, 0.053)	0.036	(0.024, 0.048)	0.053	(0.039, 0.067)
Con	Bartlett	0.045	(0.032, 0.058)	0.036	(0.024, 0.048)	0.049	(0.036, 0.062)
	Thurstone	0.045	(0.032, 0.058)	0.036	(0.024, 0.048)	0.049	(0.036, 0.062)
				S = 0.5			
Uncon	Bartlett	0.044	(0.031, 0.057)	0.051	(0.037, 0.065)	0.055	(0.041, 0.069)
	Thurstone	0.048	(0.035, 0.061)	0.047	(0.034, 0.060)	0.046	(0.033, 0.059)
	Sum Scores	0.046	(0.033, 0.059)	0.042	(0.030, 0.054)	0.051	(0.037, 0.065)
Pcon	Bartlett	0.048	(0.035, 0.061)	0.044	(0.031, 0.057)	0.049	(0.036, 0.062)
	Thurstone	0.047	(0.034, 0.060)	0.045	(0.032, 0.058)	0.049	(0.036, 0.062)
	Sum Scores	0.049	(0.036, 0.062)	0.042	(0.030, 0.054)	0.049	(0.036, 0.062)
Con	Bartlett	0.046	(0.033, 0.059)	0.045	(0.032, 0.058)	0.052	(0.038, 0.066)
	Thurstone	0.046	(0.033, 0.059)	0.045	(0.032, 0.058)	0.052	(0.038, 0.066)
				S = 0.8			
Uncon	Bartlett	0.042	(0.030, 0.054)	0.050	(0.036, 0.064)	0.054	(0.040, 0.068)
	Thurstone	0.044	(0.031, 0.057)	0.048	(0.035, 0.061)	0.047	(0.034, 0.060)
	Sum Scores	0.049	(0.036, 0.062)	0.047	(0.034, 0.060)	0.054	(0.040, 0.068)
Pcon	Bartlett	0.045	(0.032, 0.058)	0.046	(0.033, 0.059)	0.045	(0.032, 0.058)
	Thurstone	0.045	(0.032, 0.058)	0.046	(0.033, 0.059)	0.045	(0.032, 0.058)
	Sum Scores	0.042	(0.030, 0.054)	0.045	(0.032, 0.058)	0.048	(0.035, 0.061)
Con	Bartlett	0.044	(0.031, 0.057)	0.047	(0.034, 0.060)	0.045	(0.032, 0.058)
	Thurstone	0.044	(0.031, 0.057)	0.047	(0.034, 0.060)	0.045	(0.032, 0.058)

Note: Uncon, Pcon, Con denotes unconstrained, partially constrained and fully constrained latent models, respectively. N, S, and CI denotes sample size, response shift, and confidence interval resp.

Table 3: Type I Error Rates for Γ_{uneq} and α_{05}

		S = 0					
	N	*N* = 200		*N* = 500		*N* = 1000	
Model	**Factor Method**	**Rate**	**95% CI**	**Rate**	**95% CI**	**Rate**	**95% CI**
Uncon	Bartlett	0.100	(0.081, 0.119)	0.095	(0.077, 0.113)	0.097	(0.079, 0.115)
	Thurstone	0.070	(0.054, 0.086)	0.056	(0.042, 0.070)	0.058	(0.044, 0.072)
	Sum Scores	0.088	(0.070, 0.106)	0.204	(0.179, 0.229)	0.377	(0.347, 0.407)
Pcon	Bartlett	0.049	(0.036, 0.062)	0.047	(0.034, 0.060)	0.053	(0.039, 0.067)
	Thurstone	0.053	(0.039, 0.067)	0.043	(0.030, 0.056)	0.050	(0.036, 0.064)
	Sum Scores	0.045	(0.032, 0.058)	0.041	(0.029, 0.053)	0.051	(0.037, 0.065)
Con	Bartlett	0.053	(0.039, 0.067)	0.043	(0.030, 0.056)	0.050	(0.036, 0.064)
	Thurstone	0.053	(0.039, 0.067)	0.043	(0.030, 0.056)	0.050	(0.036, 0.064)
		S = 0.2					
Uncon	Bartlett	0.092	(0.074, 0.110)	0.097	(0.079, 0.115)	0.090	(0.072, 0.108)
	Thurstone	0.085	(0.068, 0.102)	0.138	(0.117, 0.159)	0.216	(0.190, 0.242)
	Sum Scores	0.229	(0.203, 0.255)	0.561	(0.530, 0.592)	0.840	(0.817, 0.863)
Pcon	Bartlett	0.222	(0.196, 0.248)	0.520	(0.489, 0.551)	0.796	(0.771, 0.821)
	Thurstone	0.215	(0.190, 0.240)	0.525	(0.494, 0.556)	0.801	(0.776, 0.826)
	Sum Scores	0.216	(0.190, 0.242)	0.506	(0.475, 0.537)	0.782	(0.756, 0.808)
Con	Bartlett	0.220	(0.194, 0.246)	0.522	(0.491, 0.553)	0.792	(0.767, 0.817)
	Thurstone	0.220	(0.194, 0.246)	0.522	(0.491, 0.553)	0.792	(0.767, 0.817)
		S = 0.5					
Uncon	Bartlett	0.093	(0.075, 0.111)	0.094	(0.076, 0.112)	0.091	(0.073, 0.109)
	Thurstone	0.195	(0.170, 0.220)	0.377	(0.347, 0.407)	0.593	(0.563, 0.623)
	Sum Scores	0.578	(0.547, 0.609)	0.920	(0.903, 0.937)	0.994	(0.989, 0.999)
Pcon	Bartlett	0.786	(0.761, 0.811)	0.991	(0.985, 0.997)	1.000	(1.000, 1.000)
	Thurstone	0.789	(0.764, 0.814)	0.991	(0.985, 0.997)	1.000	(1.000, 1.000)
	Sum Scores	0.768	(0.742, 0.794)	0.990	(0.984, 0.986)	1.000	(1.000, 1.000)
Con	Bartlett	0.783	(0.757, 0.809)	0.991	(0.985, 0.997)	1.000	(1.000, 1.000)
	Thurstone	0.783	(0.757, 0.809)	0.991	(0.985, 0.997)	1.000	(1.000, 1.000)
		S = 0.8					
Uncon	Bartlett	0.094	(0.076, 0.112)	0.095	(0.077, 0.113)	0.092	(0.074, 0.110)
	Thurstone	0.278	(0.250, 0.306)	0.550	(0.519, 0.581)	0.826	(0.803, 0.850)
	Sum Scores	0.810	(0.786, 0.834)	0.991	(0.985, 0.997)	1.000	(1.000, 1.000)
Pcon	Bartlett	0.963	(0.951, 0.975)	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)
	Thurstone	0.963	(0.951, 0.975)	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)
	Sum Scores	0.965	(0.954, 0.976)	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)
Con	Bartlett	0.966	(0.955, 0.977)	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)
	Thurstone	0.966	(0.955, 0.977)	1.000	(1.000, 1.000)	1.000	(1.000, 1.000)

Note: Uncon, Pcon, Con denotes unconstrained, partially constrained and fully constrained latent models, respectively. N, S, and CI denotes sample size, response shift, and confidence interval resp.

4.3 Power

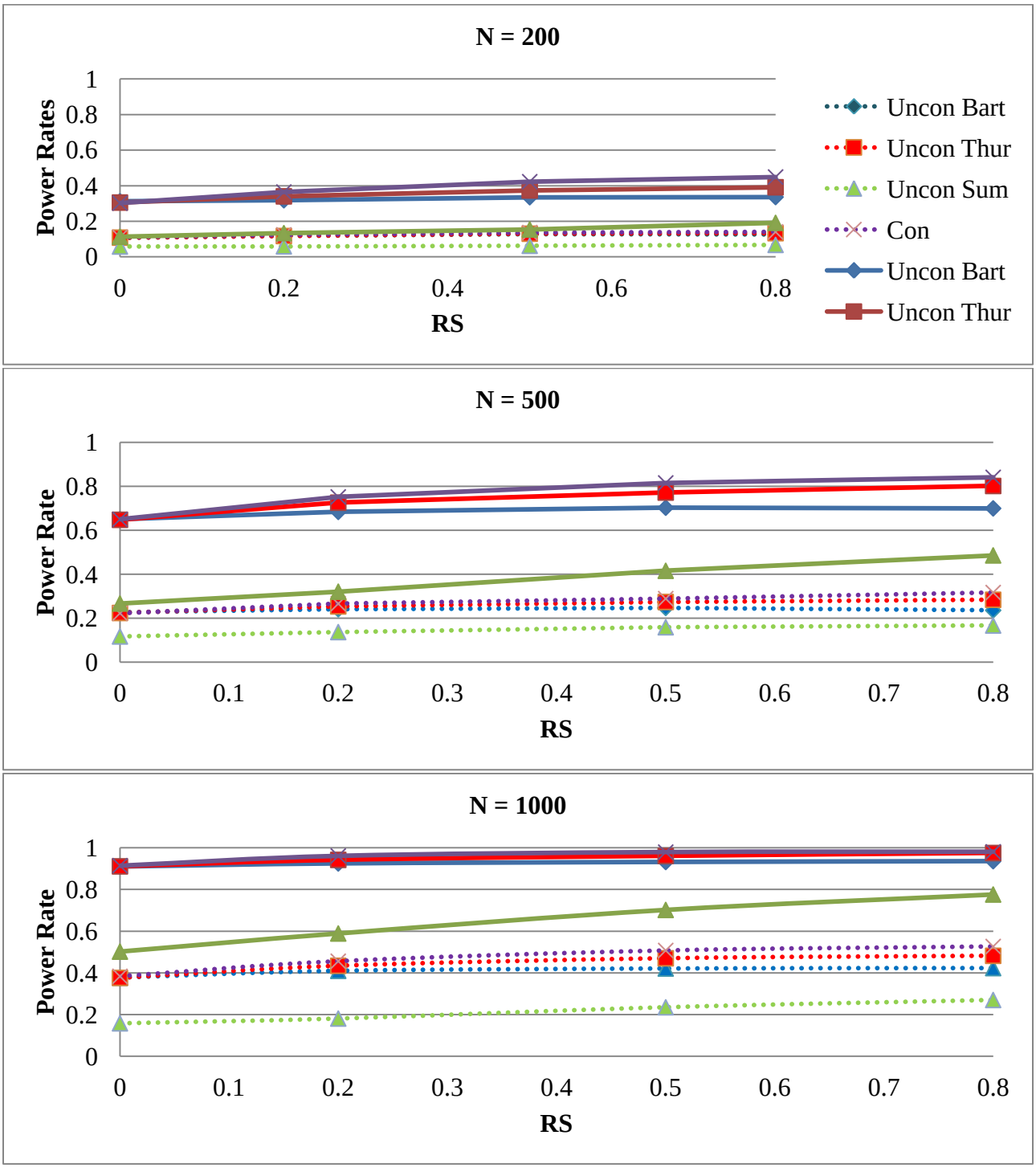
The proportion power rates (i.e. the rejection rates for $d \neq 0$) for the factor scoring methods are displayed in the graphs below (Figure 2 and 3). The power rates are the rejection rates when the null hypothesis is false. Therefore, the rejection rates were calculated for latent variable means which were unequal at time 1 and time 2. In this case $\boldsymbol{\alpha}_0$ denotes the latent variables means that were zero at time 1 and $\boldsymbol{\alpha}_{05}$ the latent variable means that were 0.5 at time 1. Again only the unequal factor loadings ($\boldsymbol{\Gamma}_{\text{uneq}}$) are displayed since the results were similar between unequal and equal factor loadings. Also, the power rates between the constrained factor scoring methods were all similar and therefore an average of all constrained factor scoring methods (Con) is displayed. Figure 2 displays the power rates for the case in which the latent variable has mean zero at time 1 ($\boldsymbol{\alpha}_0$) and Figure 3 displays power rates for the case of non-zero latent variable mean ($\boldsymbol{\alpha}_{05}$). Each figure has three separate panels for the three sample sizes (N = 200, 500, 1000).

Overall the power rates were affected by sample size, ES, RS, and latent variable mean. It is already known that the power rate increases as sample size and ES increases. The patterns of change for the power rates were similar across sample and ES.

For the case in which the latent variable mean was zero at time 1 ($\boldsymbol{\alpha}_0$) the power rates were similar for all factor scoring methods and models except, the unconstrained sum scores rates were generally smaller. From Figure 2 with sample size 200 and d = 0.05 (the solid lines) the rate for the unconstrained sum scores method was 0.06 and for all other methods it was 0.10. The unconstrained sum scores method had the lowest rates whereas the constrained method had the largest rates. Also the power rates were similar as RS increased for all methods within ES and sample size.

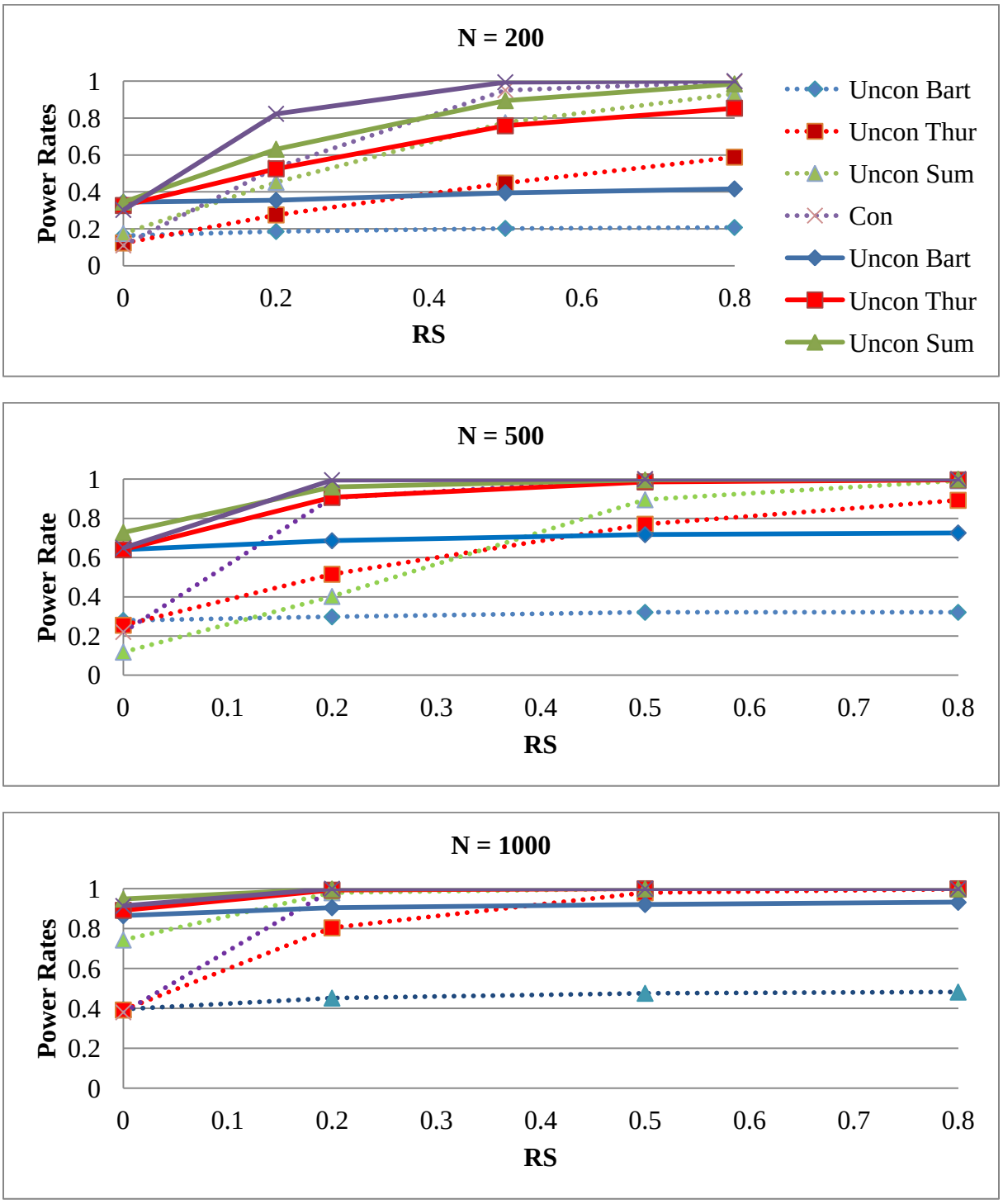
With non-zero latent variable means ($\boldsymbol{\alpha}_{05}$) the power rates differed between factor scoring methods and magnitudes of RS. When RS was absent ($S = 0$) the power rates for all the factor scoring methods were similar (with equal ES and sample size). However, as RS increased most methods increased in power rates. Only the unconstrained Bartlett method was unaffected by RS, from Figure 3 the power rates were approximately 0.20, 0.30, and 0.45 for all RS values when sample size was 200, 500, and 1000. For the other methods, since the ES is the same then the power rate increased because of the increasing RS. Therefore, RS affected the power rates; similar to the Type I error rates. Only the unconstrained Bartlett method produced factor scores with the power rates unaffected by RS regardless of sample size and latent variable mean. When the latent variable mean was zero at time 1 all factor scoring methods were unaffected by RS. The unconstrained sum scores method had the lowest power rates and would not be a good factor scoring method in this case.

Figure 2: Power Rates for α_0



Note: Solid lines represent d = 0.10 and dashed lines for d = 0.05. The diamonds denote the unconstrained Bartlett factor scores, the squares the unconstrained Thurstone, the triangles the unconstrained sum scores, and the X's denote the constrained factor scoring methods.

Figure 3: Power Rates for α_{05}



Note: Solid lines represent d = 0.05 and dashed lines for d = 0.10. The diamonds denote the unconstrained Bartlett factor scores, the squares the unconstrained Thurstone, the triangles the unconstrained sum scores, and the X's denote the constrained factor scoring methods.

4.4 Bias

Factor loadings:

The relative bias values across different models for the factor loadings are displayed in Table 4. The fully constrained (Con) and partially constrained (Pcon) models had the same relative bias regardless of sample size, ES, RS magnitude, and latent variable mean. Both models constrain the factor loadings across time and therefore produce the same factor loading estimates. Also, the relative bias values were the same for different latent variable means ($\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_{05}$), and ES values (d = 0, 0.05, and 0.10). This is because the factor loading estimates were not affected by the latent variable means. As a result, Pcon, $\boldsymbol{\alpha}$, and ES values were omitted from these tables. Since the relative biases were similar between sample sizes the table shows the average of relative biases across sample sizes.

The relative bias for the factor loadings were affected by factor loading values, model type, and RS greater than 0.2. The unconstrained model was unaffected by RS and the constrained model increased in average relative bias as RS increased. For example, from Table 3 the average relative biases (see equation 3.18) for the unconstrained model with equal factor loadings for RS values 0, 0.2, 0.5, and 0.8 were: -1.01, -0.89, -0.69, and -0.53 respectively. However, for the constrained model average relative bias was: -0.48, 0.11, 2.50, and 5.94. The results were similar for the unequal factor loadings. Also, the relative biases for equal factor loadings were smaller in magnitude than unequal factor loadings.

The relative bias for the unconstrained model always had a small negative value for the first loading and small positive values for the other loadings. This trend occured at both time points for all factor loading values and RS magnitudes. The constrained model had a similar pattern when RS was 0 or 0.2. However, when RS was greater than 0.2 the relative bias values at

time one were all positive and all negative at time two. For example, in Table 3 the relative biases for all eight loadings (four at each time point) were approximately -20, 5, 5, 5, -20, 5, 5, and 5 for both the constrained and unconstrained models with $S = 0$. When $S = 0.8$ the unconstrained model remained the same but the constrained model became 11.54, 52.98, 36.75, 24.83, -24.19, -21.26, -17.95, and -15.16. The relative biases of the factor loadings were unaffected by RS only when estimated from the unconstrained model. The unconstrained model was the better model to estimate refined factor scores when RS was present than the constrained model (at least with the factor loadings being unconstrained).

Table 4: Relative Bias Averaged across Sample Sizes

		Γ_{eq}							
	RS	0		0.2		0.5		0.8	
	Model	**Uncon**	**Con**	**Uncon**	**Con**	**Uncon**	**Con**	**Uncon**	**Con**
Time 1	γ1	-22.86	-16.87	-22.39	-9.67	-20.62	1.09	-18.55	11.54
	γ2	6.28	4.96	7.54	16.94	8.54	34.96	8.86	52.98
	γ3	6.39	5.08	6.17	12.99	5.62	24.82	5.01	36.75
	γ4	6.32	4.93	5.01	9.96	3.36	17.37	2.15	24.83
Time 2	γ1	-22.77	-16.87	-17.48	-19.20	-11.70	-21.91	-8.01	-24.19
	γ2	6.17	4.96	4.03	-5.36	2.39	-15.08	1.52	-21.26
	γ3	6.22	5.08	4.72	-3.16	3.02	-11.89	1.94	-17.95
	γ4	6.15	4.93	5.28	-1.63	3.86	-9.34	2.81	-15.16
	average	-1.01	-0.48	-0.89	0.11	-0.69	2.50	-0.53	5.94
					Γ_{uneq}				
Time 1	γ1	-28.38	-20.95	-27.36	-13.84	-24.56	-3.52	-21.65	6.13
	γ2	13.29	11.98	13.00	23.09	12.26	39.49	11.41	55.88
	γ3	11.63	8.68	11.37	17.70	10.34	31.45	9.15	45.59
	γ4	10.36	6.14	9.93	14.07	8.67	26.27	7.34	38.85
Time 2	γ1	-28.24	-20.95	-21.25	-21.74	-13.74	-22.97	-9.11	-24.41
	γ2	13.16	11.98	9.62	-0.39	6.07	-12.23	3.88	-19.76
	γ3	11.43	8.68	7.88	-1.92	4.48	-12.36	2.64	-19.12
	γ4	10.17	6.14	7.05	-3.06	4.10	-12.43	2.53	-18.66
	average	1.68	1.46	1.28	1.74	0.95	4.21	0.77	8.06

Note: Uncon, Pcon, Con denotes unconstrained, partially constrained and fully constrained latent models respectively. N denotes sample size and RS response shift. The γ_j represent the factor loading items for j=1, 2, 3, and 4.

Factor Scores:

Tables 5 to 8 contain the relative biases for factor score means for each factor scoring method. For every factor scoring method the relative bias was similar for all ES values therefore only $d = 0.10$ was reported.

The relative bias was affected by the latent variable mean. When the latent mean was zero at time 1 ($\boldsymbol{\alpha}_0$) there was little change in the bias values as RS increased for most methods. Also, the values remained similar as sample size increased and between the two types of factor loadings ($\boldsymbol{\Gamma}_{eq}$ and $\boldsymbol{\Gamma}_{uneq}$) for most methods. The only exception was the Thurstone method; the relative biases for the Thurstone method decreased in magnitude as RS increased (seen in Tables 5 and 7). Also, the relative bias for the Thurstone method was affected by the type of factor loadings. The relative biases were larger, in magnitude, when the factor loadings were unequal. Overall the relative biases for all methods were smaller for latent mean of $\boldsymbol{\alpha}_0$ than $\boldsymbol{\alpha}_{05}$.

For $\boldsymbol{\alpha}_{05}$ the relative bias was larger for the unconstrained model when there was no RS but larger for the constrained models when RS was non-zero. The absolute value of the average relative bias was generally under one for $\boldsymbol{\alpha}_0$. When the latent mean was $\boldsymbol{\alpha}_{05}$ the absolute average relative bias was almost always greater than one. The relative biases for the Thurstone method were larger than the other methods. None of the methods were affected by sample size or factor loading type. However, as RS increased there was an increase in relative bias for all methods.

The Thurstone method (for all models) had the largest relative bias. The sum scores and Bartlett method both had smaller biases and were least affected by RS.

Table 5: Factor Score Relative Bias for d=0.10, Γ_{eq}, and α_0

		S = 0								
	N	200			500			1000		
	Model	**Uncon**	**Con**	**Pcon**	**Uncon**	**Con**	**Pcon**	**Uncon**	**Con**	**Pcon**
Time 1	Bartlett	-0.16	-0.07	-0.08	-0.06	0.01	0.00	-0.05	0.00	0.00
	Thurstone	-7.24	8.15	8.26	7.10	7.61	7.64	6.94	7.33	7.33
	Sum Scores	0.04	-0.02	-0.02	0.07	0.06	0.06	0.05	0.05	0.03
Time 2	Bartlett	-0.61	-0.47	-0.47	-0.15	-0.08	-0.08	-0.21	-0.13	-0.13
	Thurstone	-7.10	-7.84	-7.93	-6.60	-6.99	-7.02	-6.43	-6.76	-6.77
	Sum Scores	-0.22	-0.19	-0.19	-0.02	0.03	0.03	0.00	0.00	-0.01
		S = 0.2								
Time 1	Bartlett	-0.13	-0.06	-0.08	-0.03	0.02	0.01	-0.02	0.02	0.02
	Thurstone	6.79	6.18	6.29	6.72	5.91	5.95	6.55	5.68	5.69
	Sum Scores	0.04	-0.02	-0.02	0.07	0.06	0.06	0.05	0.05	0.03
Time 2	Bartlett	-0.51	0.31	0.29	-0.10	0.70	0.69	-0.16	0.64	0.63
	Thurstone	-4.71	-6.69	-6.70	-4.31	-6.09	-6.09	-4.22	-5.93	-5.93
	Sum Scores	-0.22	-0.19	-0.19	-0.02	0.04	0.04	0.10	-0.01	-0.11
		S = 0.5								
Time 1	Bartlett	-0.09	-0.04	-0.08	0.00	0.04	0.03	0.00	0.03	0.03
	Thurstone	5.81	4.11	4.18	5.81	3.99	4.00	5.65	3.82	3.80
	Sum Scores	0.04	-0.02	-0.02	0.07	0.06	0.06	0.05	0.05	0.03
Time 2	Bartlett	-0.40	1.29	1.25	-0.04	1.68	1.66	-0.11	1.61	1.59
	Thurstone	-2.68	-5.40	-5.39	-2.35	-4.92	-4.94	-2.34	-4.82	-4.85
	Sum Scores	-0.23	-0.21	-0.21	-0.01	0.04	0.04	0.25	-0.01	-0.27
		S = 0.8								
Time 1	Bartlett	-0.06	-0.02	-0.06	0.02	0.06	0.05	0.02	0.05	0.05
	Thurstone	4.89	2.74	2.75	4.93	2.69	2.65	4.79	2.55	2.49
	Sum Scores	0.04	-0.02	-0.02	0.07	0.06	0.06	0.05	0.05	0.03
Time 2	Bartlett	-0.32	2.10	2.04	-0.01	2.48	2.45	-0.08	2.40	2.37
	Thurstone	-1.69	-4.35	-4.36	-1.40	-3.93	-3.98	-1.42	-3.87	-3.94
	Sum Scores	-0.24	-0.22	-0.22	0.00	0.05	0.05	0.40	-0.01	-0.43

Note:Uncon, Pcon, Con, N, S, d and CI denotes unconstrained, partially constrained and fully constrained latent models, sample size, response shift, effect size, and confidence interval respectively.

Table 6: Factor Score Relative Bias for d=0.10, Γ_{eq}, and α_{05}

		S = 0								
	N	200			500			1000		
	Model	**Uncon**	**Con**	**Pcon**	**Uncon**	**Con**	**Pcon**	**Uncon**	**Con**	**Pcon**
Time 1	Bartlett	-5.97	-1.71	-3.45	-3.71	-1.02	-2.04	-3.02	-0.86	-1.71
	Thurstone	26.54	10.56	31.48	27.42	10.22	30.07	27.23	9.97	29.24
	Sum Scores	-0.06	-0.02	0.12	0.19	0.06	0.25	0.17	0.05	0.10
Time 2	Bartlett	-5.52	-2.11	-3.52	-3.12	-1.10	-1.84	-2.80	-0.99	-1.64
	Thurstone	-13.72	-5.43	-13.96	-11.12	-4.38	-11.21	-10.32	-4.12	-10.52
	Sum Scores	-0.52	-0.19	-0.61	0.09	0.03	-0.05	-0.01	0.00	-0.03
					S = 0.2					
Time 1	Bartlett	-5.15	-5.07	-10.19	-3.21	-4.53	-9.11	-2.64	-4.42	-8.90
	Thurstone	28.41	11.56	32.99	29.11	11.18	31.47	28.77	10.85	30.38
	Sum Scores	-0.06	-0.02	0.12	0.19	0.06	0.25	0.17	0.05	0.10
Time 2	Bartlett	-4.34	2.70	4.32	-2.39	3.59	5.88	-2.21	3.65	6.01
	Thurstone	-10.21	-7.95	-18.57	-8.26	-6.91	-16.20	-7.74	-6.59	-15.49
	Sum Scores	1.04	-0.19	-2.03	1.60	0.04	-1.52	1.50	-0.01	-1.52
					S = 0.5					
Time 1	Bartlett	-4.12	-9.29	-18.89	-2.56	-8.92	-18.14	-2.12	-12.47	-18.05
	Thurstone	26.66	9.86	26.33	27.23	9.49	24.85	26.83	6.42	23.70
	Sum Scores	-0.06	-0.02	0.12	0.19	0.06	0.25	0.17	0.05	0.10
Time 2	Bartlett	-3.03	8.68	13.96	-1.58	9.42	15.36	-1.53	14.02	15.40
	Thurstone	-5.69	-7.92	-16.94	-4.32	-6.96	-15.09	-4.11	-4.89	-14.53
	Sum Scores	2.76	-0.21	-3.55	3.27	0.04	-3.10	3.16	-0.01	-3.11
					S = 0.8					
Time 1	Bartlett	-3.35	-12.74	-26.10	-2.04	-12.48	-25.58	-1.69	-12.48	-25.55
	Thurstone	23.45	7.11	17.68	23.94	6.77	16.36	23.55	6.77	15.29
	Sum Scores	-0.06	-0.02	0.12	0.19	0.06	0.25	0.17	0.06	0.10
Time 2	Bartlett	-2.17	13.46	21.69	-1.05	14.09	22.95	-1.08	14.09	22.92
	Thurstone	-3.17	-6.02	-12.38	-2.13	-5.16	-10.91	-2.10	-5.16	-10.50
	Sum Scores	4.02	-0.22	-4.63	4.49	0.05	-4.22	4.38	0.05	-4.24

Note:Uncon, Pcon, Con, N, S, d and CI denotes unconstrained, partially constrained and fully constrained latent models, sample size, response shift, effect size, and confidence interval respectively.

Table 7: Factor Score Relative Bias for d=0.10, Γ_{uneq}, and α_0

		S =0								
	N	200			500			1000		
	Model	**Uncon**	**Con**	**Pcon**	**Uncon**	**Con**	**Pcon**	**Uncon**	**Con**	**Pcon**
Time 1	Bartlett	-0.27	-0.16	-0.17	-0.10	-0.01	-0.02	-0.11	-0.03	-0.03
	Thurstone	11.01	12.17	13.05	9.45	10.05	10.10	9.02	9.49	9.50
	Sum Scores	0.04	-0.02	-0.02	0.07	0.05	0.05	0.02	0.05	0.05
Time 2	Bartlett	-0.78	-0.63	-0.63	-0.26	-0.14	-0.14	-0.34	-0.21	-0.21
	Thurstone	-10.92	-11.94	-12.88	-9.01	-9.40	-9.45	-8.55	-8.91	-8.93
	Sum Scores	-0.22	-0.18	-0.18	-0.02	0.03	0.03	-0.01	0.00	0.00
		S = 0.2								
Time 1	Bartlett	-0.22	-0.13	-0.15	-0.07	0.01	0.00	-0.08	-0.01	-0.01
	Thurstone	9.42	8.29	8.48	8.72	7.59	7.65	8.38	7.21	7.23
	Sum Scores	0.04	-0.02	-0.02	0.07	0.05	0.05	0.02	0.05	0.05
Time 2	Bartlett	-0.65	0.25	0.24	-0.19	0.72	0.72	-0.27	0.64	0.64
	Thurstone	-6.36	-8.92	-8.97	-5.48	-7.79	-7.80	-5.28	-7.49	-7.49
	Sum Scores	-0.22	-0.19	-0.19	-0.02	0.03	0.03	-0.01	-0.01	-0.01
		S = 0.5								
Time 1	Bartlett	-0.17	-0.10	-0.11	-0.03	0.04	0.03	-0.05	0.02	0.02
	Thurstone	7.50	5.30	5.42	7.32	5.07	5.10	7.06	4.81	4.81
	Sum Scores	0.04	-0.02	-0.02	0.07	0.05	0.05	0.02	0.05	0.05
Time 2	Bartlett	-0.49	1.34	1.32	-0.10	1.79	1.79	-0.18	1.70	1.71
	Thurstone	-3.19	-6.75	-6.74	-2.78	-6.09	-6.10	-2.74	-5.92	-5.94
	Sum Scores	-0.23	-0.20	-0.20	-0.01	0.04	0.04	-0.01	-0.01	-0.01
		S = 0.8								
Time 1	Bartlett	-0.14	-0.06	-0.08	-0.01	0.06	0.05	-0.03	0.04	0.04
	Thurstone	6.15	3.54	3.60	6.09	3.44	3.43	5.89	3.25	3.21
	Sum Scores	0.04	-0.02	-0.02	0.07	0.05	0.05	0.02	0.05	0.05
Time 2	Bartlett	-0.39	2.22	2.19	-0.04	2.64	2.64	-0.13	2.55	2.55
	Thurstone	-1.91	-5.35	-5.34	-1.58	-4.82	-4.85	-1.60	-4.73	-4.77
	Sum Scores	-0.24	-0.21	-0.21	0.00	0.05	0.05	-0.02	-0.01	-0.01

Note:Uncon, Pcon, Con, N, S, d and CI denotes unconstrained, partially constrained and fully constrained latent models, sample size, response shift, effect size, and confidence interval respectively.

Table 8: Factor Score Relative Bias for d=0.10, Γ_{uneq}, and α_{05}

		S = 0							
	N	200			500			1000	
	Model	**Uncon**	**Con**	**Pcon**	**Uncon**	**Con**	**Pcon**	**Uncon**	**Con**	**Pcon**
Time 1	Bartlett	-3.96	-2.58	-2.59	-2.96	-1.80	-1.80	-2.69	-1.66	-1.66
	Thurstone	12.40	14.52	15.46	11.42	12.60	12.68	11.08	12.08	12.12
	Sum Scores	0.04	-0.02	-0.02	0.07	0.05	0.05	0.02	0.05	0.05
Time 2	Bartlett	-4.36	-3.04	-3.04	-3.04	-1.93	-1.94	-2.93	-1.84	-1.83
	Thurstone	-9.42	-9.60	-10.38	-7.10	-6.85	-6.86	-6.47	-6.31	-6.32
	Sum Scores	-0.22	-0.18	-0.18	-0.02	0.03	0.03	-0.01	0.00	0.00
						S = 0.2				
Time 1	Bartlett	-3.42	-6.10	-6.11	-2.56	-5.51	-5.52	-2.36	-5.43	-5.44
	Thurstone	13.83	15.67	16.13	13.21	14.51	14.65	12.84	13.92	13.99
	Sum Scores	0.04	-0.02	-0.02	0.07	0.05	0.05	0.02	0.05	0.05
Time 2	Bartlett	-3.41	2.33	2.27	-2.34	3.30	3.27	-2.30	3.33	3.34
	Thurstone	-6.93	-12.18	-12.23	-5.46	-10.26	-10.26	-5.09	-9.69	-9.70
	Sum Scores	-0.22	-0.19	-0.19	-0.02	0.03	0.03	-0.01	-0.01	-0.01
						S = 0.5				
Time 1	Bartlett	-2.69	-10.43	-10.52	-2.00	-10.04	-10.13	-1.86	-10.01	-10.10
	Thurstone	13.31	13.52	13.87	13.19	12.77	12.85	12.86	12.21	12.19
	Sum Scores	0.04	-0.02	-0.02	0.07	0.05	0.05	0.02	0.05	0.05
Time 2	Bartlett	-2.32	8.92	8.77	-1.51	9.71	9.65	-1.53	9.68	9.67
	Thurstone	-3.68	-11.76	-11.72	-2.93	-10.34	-10.38	-2.79	-9.84	-9.93
	Sum Scores	-0.23	-0.20	-0.20	-0.01	0.04	0.04	-0.01	-0.01	-0.01
						S = 0.8				
Time 1	Bartlett	-2.11	-13.92	-14.12	-1.52	-13.65	-13.85	-1.42	-13.65	-13.86
	Thurstone	12.08	10.26	10.40	12.07	9.66	9.57	11.79	9.15	8.95
	Sum Scores	0.04	-0.02	-0.02	0.07	0.05	0.05	0.02	0.05	0.05
Time 2	Bartlett	-1.60	14.12	13.88	-0.96	14.78	14.67	-1.02	14.70	14.64
	Thurstone	-2.01	-9.41	-9.46	-1.46	-8.23	-8.41	-1.44	-7.82	-8.06
	Sum Scores	-0.24	-0.21	-0.21	0.00	0.05	0.05	-0.02	-0.01	-0.01

Note:Uncon, Pcon, Con, N, S, d and CI denotes unconstrained, partially constrained and fully constrained latent models, sample size, response shift, effect size, and confidence interval respectively.

Chapter 5: Discussion

5.1 Conclusion

PROs are becoming more popular in clinical and epidemiological studies[31]. However, these studies may be affected by RS in longitudinal data. RS is a type of measurement bias which can influence the inference of true change in measures over time[31]. Subject-specific refined factor scores are estimated from latent variable models and can be used in further analyses. The most common factor scoring methods include the Thurstone[53], Bartlett[4], and sum scores method[29, 30]. It is uncertain how factor score estimation will be affected by RS. This study was done to determine which factor scoring method is least affected by the presence of reprioritization RS.

A simulation study was conducted to compare the effects of RS on factor scoring methods. The study used three CFA models to analyze data from two time points. The models had different set of parameters invariant over time and were named unconstrained, fully constrained and partially constrained. Data was generated with different magnitudes of reprioritization RS. Factor scores were estimated and change of the factor scores means were tested by a paired t-test. The performance of each factor scoring method was determined by measuring Type I error, statistical power, and relative bias for the factor loadings and factor scores. These measures were compared across magnitudes of RS to inspect which factor score estimates are least affected by reprioritization RS.

The Type I error rate was affected by RS, sample size, model type, factor score method, and latent variable mean. With the latent variable means equal to zero the rejection rates were approximately 0.05 for all factor scoring methods regardless of sample size, model type, and RS.

With the latent means equal to 0.5 at both times the rejection rates increased as sample size and RS increased. The unconstrained Bartlett method is the only exception; the rejection rates stayed at approximately 0.10 consistently. Overall the unconstrained model rejection rates were least affected by increasing RS magnitude.

The power rates were affected by sample size, ES, RS, latent variable mean, and model type. From previous studies it is known that power rates are affected by sample size and ES; as sample size and ES increase the power rate increases. As RS increased the rejection rate also increased but was influenced by the latent variable mean. With a zero latent variable mean (at time 1) the rejection rate increased only slightly (i.e. 0.12 to 0.14 for $N = 200$ and $d = 0.05$) as opposed to non-zero latent variable mean (at time 1) where the increase in rate was more considerable (i.e 0.11 to 0.99 for the constrained models, $d = 0.05$, and $N = 200$). However, as with the Type I error rate, the unconstrained Bartlett method was less affected by RS than the other methods, for $\boldsymbol{\alpha}_{05}$ when $N = 200$ the rate increased from 0.164 to 0.205. Also, although the power rates were similar between factor scoring methods when the latent variable mean was zero at time 1 the sum scores method always had the smallest power rates for all sample sizes.

The relative bias for the factor loadings were affected by RS, model type, and factor loading values. The latent variable means and ES did not affect the factor loading estimation and therefore did not affect the relative bias. Also since both the fully constrained and partially constrained models have equality of factor loadings over time then these models had the same relative bias values. As the RS increased for the unconstrained model the average bias decreased in magnitude. Furthermore, with regards to the magnitude of relative bias, the values at time point two were less than that of time point one. Also, the first factor loading was always the largest and negative. For the constrained (and partially constrained) the average bias increased in

magnitude as RS increased. The relative bias at both time points increased in value and magnitude respectively. The relative biases at time point two for the constrained model were generally all negative. Overall the unconstrained model had smaller average relative bias (absolute value) than the constrained model when RS was present.

The factor score relative bias was influenced by certain parameters depending on the latent variable mean, and the factor scoring method. When the latent variable mean was zero the relative bias was similar across methods, sample size and models. There was a slight decrease as RS increased. The only exception is the Thurstone method which had to largest relative bias that decreased in magnitude as RS increased regardless of the model used.

However, when the latent variable mean was non-zero there were differences in bias among the factor scoring methods. As sample size increased the relative bias for the Bartlett and sum scores method decreased in magnitude but had no effect on the Thurstone method. All methods had an increase in relative bias when RS increased however; the increase was small in the constrained Thurstone and sum scores and the unconstrained Bartlett method. Overall the Thurstone method at time one had larger relative bias values and the bias values at time two were usually smaller in magnitude. Since the sum scores method was based on the sample data the relative bias at time one was not affected by RS or ES. The sum scores method at time two was affected by RS and increased slightly. Lastly, the value of the factor loadings did not affect the relative bias for the factor scores.

From this study, the recommendation of the most appropriate factor scoring method to estimate factor scores in the presence of reprioritization RS is the unconstrained Bartlett method. When the latent variable means were zero all methods performed very similarly and were not largely affect by RS. The Type I error rates were approximately 0.05 across all methods. The

power rates were similar between methods and stayed the same between different RS magnitudes. However, the sum scores method had the smallest power rates for all sample sizes. The average factor loading bias for the unconstrained model remained the same as RS increased and the constrained models increased with increasing RS. Lastly, the Thurstone factor scores had the largest bias values. From these results the unconstrained Bartlett method was not affected by RS but also performed well in all measures (Type I error rates, statistical power, factor loading and factor score bias).

For non-zero latent variable means (regardless if the means were equal across time) the unconstrained Bartlett method was less affected by RS and had consistently small relative bias. Only the unconstrained Bartlett factor scores rejection rates (for both power and Type I error) remained consistent across different magnitudes of RS present. The relative bias of the factor scores and factor loadings for unconstrained model were not affected by an increase in RS magnitude, unlike the constrained models. Also, the relative biases for the unconstrained Bartlett factor scores were less affected by RS. However, the sum scores method had lower bias values, and the unconstrained Bartlett Type I error rates were large and power rates were small. Nonetheless, since the unconstrained Bartlett factor score parameters (Type I error, power, and relative biases) remained stable as RS increased regardless of sample size, factor loading, and latent variable mean values then this is the most appropriate factor scoring method to use when the latent variable mean is non-zero.

5.2 Significance and Limitations

Significance

There are several methods that have been developed to identify RS in datasets[31, 35, 42, 48], particularly in PROs but; once RS has been identified what can be done with the data? There has been little research on the effect of RS (or measurement non-invariance in general) on the estimation of factor scores[59]. Factor scores are used in many studies such as clinical, epidemiology, psychology, and economics; many of these studies rely on PROs which can be affected by RS in longitudinal data[31]. Incorrect conclusions could be inferred if RS is ignored when estimating factor scores[47] to use in additional analyses or as a summary variable.

RS can affect the interpretation of change in measures[31], impede the comparison of repeated measures[39], could attenuate or exaggerate estimates[48], and therefore plays an important role in PROs analysis[51]. This study is the next step of RS research by determining how RS affects factor score estimation and inference from factor scores. By knowing how RS affects subsequent data analysis then research can begin on how to correct for RS.

Strengths

This study builds on the research done by Oort to identify RS[42], which has been used in other studies as a way to identify and examine the magnitude of RS in continuous data[17]. Oort's method uses an SEM model which is easy to implement and well developed[42].

Reprioritization RS is used because it affects the factor loadings[42] and thus likely to affect factor scores. It was easy to generate data with reprioritization RS and it has been examined in previous studies[31]. The refined factor score methods (Bartlett[4] and Thurstone[53]) include the

factor loading estimates in their weighting matrix[30] and therefore since RS affects the factor loading estimates it should also affect the factor scoring methods.

The parameter values were chosen from previous simulation studies[25, 29, 49]and to ensure that the residual and latent covariance matrices were positive definite. Only one latent variable was implemented over time. This was for simplicity, to ensure there was no confusion over results based on RS or because of correlation between multiple latent variables or variables loading onto more than one latent variable.

Limitations

Modeling with estimated factor scores has the potential to lead to biased conclusions about the latent structural relations. Factor scores, in general, are sensitive to the factor extraction method and rotation method (i.e. methods to calculate the factor loadings) in EFA or CFA.

Another consideration when creating factor scores is indeterminacy, or lack of uniqueness of factor scores. Indeterminacy arises under the common factor model, because the parameter estimates determined by the researcher's choice of the communality estimate (i.e. the estimate of the proportion of the variance of the variable that is both error free and shared with other variables in the matrix). This means that there is not a unique solution for the factor analysis results and, theoretically, an infinite number of solutions could account for the relationships between items and factor(s) (i.e. $\mathbf{\Gamma}$). Therefore, it also follows that the factor scores are not uniquely defined[13, 21]. However, Lawley and Maxwell[30] developed a method to choose a factor loading matrix, $\mathbf{\Gamma}$, by solving $\mathbf{J} = \mathbf{\Gamma}^{T}[\text{diag}(\mathbf{\Theta})]\mathbf{\Gamma}$, where $\mathbf{J}$ is a diagonal matrix, and ^{T} is the transpose operator $\mathbf{\Gamma}$ can be estimated. The possibility of any diagonal elements of $\mathbf{J}$ being equal is ignored (since it is unlikely); this fixes $\mathbf{\Gamma}$ except that any column may have its elements reversed in sign. This method of $\mathbf{\Gamma}$ estimation is the convention in factor analysis and is

implemented by PROC CALIS. Factor score indeterminacy is a source of controversy. Since it is possible to have an infinite number of solutions it is hard to know if the conclusions from estimated factor scores are reliable. The degree of indeterminacy can be examined by assessing how strongly the estimated factor scores correlate to their respective factors.

Another limitation of the study is the generalizability of the simulation results. The simulated data had no missing observations as the effects of missing data are still being researched for PROs[5] let alone for PROs affected with RS. Longitudinal PROs frequently have missing data due to patient attrition and non-response on individual items[5]. The amount of missing data in longitudinal PROs can be very large. Missing data reduces the number of observations available for analysis; this can cause problems since latent variable models require large sample sizes[38]. The simulated data only had two time points since most of the RS research has not extended past two time points[31, 35, 42, 47, 48]. Also, the data was continuous and normally distributed. Lastly, the model contained only a single latent variable at each time point and an equal number of items loading onto each latent variable.

Future Studies

Future work could include applying this method to a real data set and using more complex data such as, more than one latent variable, cross loadings of response items, or unequal observed variables loading on each latent variable. Also there could be studies on factor score estimation from discrete variables or a mixture of discrete and continuous variables[25, 49]. This study only examined three factor scoring methods however there are other factor scoring methods that could be studied[25, 29, 49]. Lastly, the SEM method is a group level RS method. Since factor scores are individual level then individual level RS methods may result in different conclusions[35].

References

[1] D.L. Bandalos. The Effects of Item Parceling on Goodness-of-fit and Parameter Estimate Bias in Structural Equation Modeling. *Structural Equation Modeling,* 9(2002):78–102.

[2] R. Barclay-Goddard, L. Lix, R. Tate, L. Weinberg, N. E. Mayo. Does Response Shift Occur in Self-Perceived Physical Function after Stroke? *Archives of Physical Medicine and Rehabilitation,* 92(2011):1762-1769.

[3] D.J. Bartholomew. Posterior Analysis of the Factor Model. *Bristish Journal of Mathematical and Statistical Psychology,* 34(1981):93-99.

[4] M.S. Bartlett. The Statistical Conception of Mental Factors. *British Journal of Psychology,* 28(1937):97-104.

[5] M. Blanchin, J. Hardouin, T.L. Neel, G. Kubis, C. Blanchard, E. Miraillie, V. Sebille. Comparison of CTT and Rasch-based Approaches for the Analysis of Longitudinal Patient Reported Outcomes. *Statistics in Medicine,* 30(2011):825-838.

[6] K.A. Bollen. *Structural Equations with Latent Variables*. New York, NY: Wiley (1989).

[7] D.I. Boomsma, P.C.M. Molenaar, J.F. Orlebek. Estimation of Individual Genetic and Environmental Factor Scores. *Genetic Epidemiology,* 7(1990):83-91.

[8] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed). Hillsdale, NJ: Lawrence Erlbaum Associates (1988).

[9] A.L. Comrey, H.B. Lee. *A First Course in Factor Analysis* (2nd ed). Hillsdale, NJ: Erlbaum (1992).

[10] P.J. Curran, M.E. Edwards, R.J. Wirth, A.M. Hussong, L. Chassin. Alternative Categorical Measurement Models for the Analysis of Individual Growth. To appear in T. Little (Ed.), *Modeling Developmental Processes in Ecological Context*. Mahwah, NJ: Laurence Erlbaum Associates (2007).

[11] P.J. Diggle, P. Heagerty, K. Liang, S.L. Zeger. *Analysis of Longitudinal Data* (2nd ed). Cary, NC: Oxford (2002).

[12] T. Dijkstra. Some Comments on Maximum Likelihood and Partial Least Squares Methods. *Journal of Econometrics,* 22(1983):67–90.

[13] C. DiStefano, M. Zhu, D. Mindrila. Understanding and Using Factor Scores: Considerations for Applied Researcher. *Practical Assessment, Research and Evaluation,* 14(2009).

[14] D.L. Fairclough. Patient Reported Outcomes as Endpoints in Medical Research. *Statistical Methods in Medical Research,* 13(2004):115-138.

[15] P. Fayers, D. Machin. *Quality of Life: The Assessment, Analysis, and Interpretation of Patient-Reported Outcomes* (2nd ed). West Sussex: Wiley (2007).

[16] J.F. Fries, B. Bruce, D. Cella. The Promise of PROMIS: Using Item Response Theory to Improve Assessment of Patient-Reported Outcomes. *Clinical Experimental Rheumatology,* 23 (2005):53-57.

[17] P. Gandhi, L. Ried, I. Huang, C. Kimberlin, T. Kauf. Assessment of Response Shift using Two Structural Equation Modelling Techniques. *Quality of Life Research,* 22(2013):461-471.

[18] G.V. Glass, T.O. Maguire. Abuses of Factor Scores. *American Education Research Journal,* 3(1966):297-304.

[19] G. Goldstein, J.F. Luther, G.L. Haas, C.J. Appelt, A.J. Gordon. Factor Structure and Risk Factors for the Health Status of Homeless Veterans. *Psychiatric Quarterly*, 81(2010):311-323.

[20] R.L. Gorsuch. *Factor Analysis* (2nd ed). Hillsdale, NJ: Erlbaum (1983).

[21] J.W. Grice. Computing and Evaluating Factor Scores. *Psychological Methods,* 6(2001):430-450.

[22] J.W. Grice, R.J. Harris. A Comparison of Regression and Loading Weights for the Computation of Factor Scores. *Multivariate Behavioral Research,* 33(1998):221-247.

[23] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, R.L. Tatham. *Multivariate Data Analysis* (6th ed). Upper Saddle River, NJ: Pearson Prentice Hall (2006).

[24] S.L. Hershberger. Factor Scores. In B.S. Everitt and D.C. Howell (Ed.), *Encyclopedia of Statistics in Behavioral Science.* New York, NY: John Wiley (2005):636-644.

[25] T. Hoshino, P. Bentler. Bias in Factor Score Regression and a Simple Solution. In A. Leon and K. Carrier (Ed.), *Analysis of Mixed Data: Methods and Applications.* Berkley, CA: Chapman & Hall/CRC (2013):43-62.

[26] K.G Joreskog, D. Sorbom. *Lisrel7: A guide to the Program and Applications*. Chicago, IL: SPSS (1989).

[27] K.G. Joreskog, H. Wold. The ML and PLS Techniques for Modeling with Latent Variables: Historical and Comparative Aspects. *Systems under Indirect Observation: Causality, Structure, Prediction. Part II.* Amsterdam: North-Holland (1982):263-270.

[28] T. Kline. *Psychological Testing: A Practical Approach to Design and Evaluation* (1st ed). Thousand Oaks, CA:Sage (2005).

[29] J.L. Lastovicka, K. Thamodaran. Common Factor Score Estimates in Multiple Regression Problems. *Journal of Marketing Research,* 28(1991):105-112.

[30] D.N. Lawley, A.E. Maxwell. Factor Analysis as a Statistical Method. *Journal of Royal Statistical Society,* 12(1971):209-229.

[31] L.M. Lix, T. Sajobi, R. Sawatzky, J. Liu, N.E. Mayo, Y. Huang, L.A. Graff, J.R. Walker, J. Ediger, I. Clara, K. Sexton, R. Carr, C.N. Bernstein. Relative Importance Measures for Reprioritization Response Shift**.** *Quality of Life Research,* 22(2013):695-703.

[32] R.C. MacCallum, K.F. Widaman, S.Zhang, S. Hong. Sample Size in Factor Analysis. *Psychological Methods,* 4(1999):84–99.

[33] H.W. Marsh, K.T. Hau, J.R. Balla, D. Grayson. Is More Ever Too Much?: The Number of Indicators per Factors in Confirmatory Factor Analysis. *Multivariate Behavioral Research,* 33(1998):181–222.

[34] L. Mathiesen, M.H. Anderson, P.K. Hol, P.S. Lingass, R. Lundblad, K.A. Rein, T.I. Tonnessen, B.E. Mork, J. Svennevig, A.K. Wahl, B.R. Hanestad, E. Fosse. Patient-Reported Outcome After Randomization to On-Pump Versus Off-Pump Coronary Artery Surgery. *Annals of Thoracic Surgery,* 79(2005):1584-1589.

[35] N.E. Mayo, S.C. Scott, N. Dendukuri, S.Wood-Dauphinee. Identifying Response Shift Statistically at the Individual Level. *Quality of Life Research,* 17(2008):627-639.

[36] R.P. McDonald. Path Analysis with Composite Variables. *Multivariate Behavioral Research,* 33(1996):239-270.

[37] W. Meredith. Measurement Invariance, Factor Analysis and Factorial Invariance. *Psychometrika,* 58(1993):525-543.

[38] B.O. Muthén, L. K. Muthén. *MPLUS Statistical Analysis With Latent Variables: User's Guide* (6th ed). Los Angeles, CA: Muthén & Muthén (1998-2010).

[*39]* M.R. Novick. The Axioms and Principal Results of Classical Test Theory. *Journal of Mathematical Psychology,* 3(1966):1-18.

[40] T. Oga, K. Nishimura, M. Tsukino, S. Sato, T. Hajiro, M. Mishima. Longitudinal Deteriorations in Patient Reported Outcomes in Patients with COPD. *Respiratory Medicine,* 101(2007):146-153.

[41] H. Ogasawara. Standard Errors for Rotation Matrices with an Application to the Promax Solution. *British Journal of Mathematical and Statistical Psychology,* 51(1998):163–178.

[42] F.J. Oort. Using Structural Equation Modeling to Detect Response Shifts and True Change. *Quality of Life Research,* 14(2005):587-598.

[43] F.J. Oort, M.R.M. Visser, M.A.G. Sprangers. Formal Definitions of Measurement Bias and Explanation Bias Clarify Measurement and Conceptual Perspectives on Response Shift. *Journal of Clinical Epidemiology,* 62(2009):1126-1137.

[44] N. Ram, S. Chow, R. Bowles, L. Wang, K. Grimm, F. Fujita. Examining Interindividual Differences in Cyclicity of Pleasant and Unpleasant affects using Spectral Analysis and Item Response Modeling. *Psychometrika,* 70(2005):773-790.

[45] A.P. Sage, J.L. Melsa. *Estimation Theory with Applications to Communications and Control*. New York, NY: McGraw-Hill Book Company (1971).

[46] SAS Institute Inc. *SAS/STAT User's Guide, Version 9.2*. Cary, NC: SAS Institute Inc. (2008).

[47] R. Sawatzky, P.A. Ratner, J.A. Kopec, B.D. Zumbo. Latent Variable Mixture Models: A Promising Approach for the Validation of Patient Reported Outcomes. *Quality of Life Research,* 21(2011):637-650.

[48] C.E. Schwartz, M.A.G Sprangers. Methodological Approaches for Assessing Response Shift in Longitudinal Health-related Quality-of-life Research. *Social Science and Medicine*, 48(1999):1531–1548.

[49] A. Skrondal, P. Laake. Regression among Factor Scores. *Psychometrika,* 66(2001):563-576.

[50] A. Skrondal, S. Rabe-Hesketh. *Interdisciplinary Statistics Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Berkley, CA: Chapman & Hall/CRC (2004).

[51] M.A. Sprangers, C.E. Schwartz. Integrating Response Shift into Health-Related Quality of Life Research: a Theoretical Model. *Social Science & Medicine,* 48(1999):1507-1515.

[52] D.R. Thomas, I.R.R. Lu, M. Cedzynski. Partial Least Squares: A Critical Review and Potential Alternative. *Proceedings of the Administrative Sciences Association of Canada (ASAC) Conference,* (2005).

[53] L.L Thurstone. *The Vectors of the Mind*. Chicago, IL: University of Chicago Press (1935).

[54] V.E. Vinzi, C. Lauro. Partial least squares. *Computational Statistics and Data Analysis,* 48(2005):159-205.

[55] J.E. Ware, C.D. Sherbourne. The MOS 36-item Short-form Health Survey (SF-36). Conceptual Framework and Item Selection. *Medical Care,* 30(1992):473–483.

[56] K.F. Widaman, E. Ferrer, R.D. Conger. Factorial Invariance within Longitudinal Structural Equation Models: Measuring the Same Construct across Time. *Child Development Perspective,* 4 (2010):10-18.

[57] H. Wold. Non-linear Iterative Partial Least Squares (NIPALS) Modeling. Some current developments. In P. R. Krishnaiah (Ed.), *Multivariate Analysis, vol. III*. New York, NY: Academic Press (1973):383-407.

[58] W. Wu, C. Kao, L. Yen, T.S. Lee. Comparison of Children's Self-reports of Depression Symptoms among Different Family Interaction Types in Northern Taiwan. *BMC Public Health*, 7(2007):116.

[59] A.D. Wu, Z. Li, B.D. Zumbo. Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data. *Practical Assessment Research and Evaluation*, 12(2007).

Appendix: Derivations

Table A1: Statistical Derivations

Parameter	True	Bartlett[4]	Thurstone[53]
mean of f_1	α_1	α_1	$\Psi_{11}\mathbf{\Gamma}_1^{T}\mathbf{\Sigma}_1^{-1}\alpha_1$
mean of f_2	α_2	α_2	$\Psi_{22}\mathbf{\Gamma}_2^{T}\mathbf{\Sigma}_2^{-1}\alpha_2$
Regression Coefficient	$\Psi_{12}\Psi_{22}^{-1}$	$\Psi_{12}(\Psi_{22}+(\mathbf{\Gamma}_2^{T}\mathbf{\Theta}_{22}^{-1}\mathbf{\Gamma}_2)^{-1})^{-1}$	$\Psi_{11}(\mathbf{\Gamma}_1^{T}\mathbf{\Sigma}_1^{-1}\mathbf{\Gamma}_1)\,\Psi_{12}\Psi_{22}^{-1}$
Covariance Matrix	Ψ_{12}	Ψ_{12}	$\Psi_{11}\mathbf{\Gamma}_1^{T}\Sigma_1^{-1}\mathbf{\Gamma}_1\Psi_{12}\mathbf{\Gamma}_2^{T}\Sigma_2^{-1}\mathbf{\Gamma}_2\,\Psi_{22}^{-1}$

Where f_t is a factor at time t (t=1,2); $\Psi_{tt'}$ is the covariance matrix of factor t and t' ($t \neq t'$); $\mathbf{\Gamma}_t$ is the factor loading matrix for factor t; $\mathbf{\Gamma}_{tt}$ is the covariance matrix of the error term for factor t; $\mathbf{\Sigma}_t$ is the covariance matrix of $\mathbf{Y}_{it}$ for each individual i=1,…,N and factor t.